Strategic Selection Around Kindergarten Recommendations

Michael David Ricks*

This Draft Updated: August 2, 2022
For latest draft click here

Abstract

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

*Department of Economics, University of Michigan. Email: ricksmi@umich.edu.
This research is the product of generous amounts of feedback and conversations with many wonderful and intelligent people including Amanda Kowalski, Brian Jacob, Jim Hines, Sara Heller, Ash Craig, Jeff Denning, John Bonney, Chirs Walters, Rich Patterson, Joe Price, Jeff Smith, Christina Weiland, Lars Lefgren, Matias Cattaneo, Peter Hull, Jason Baron, Charlie Brown, John Bound, Mike Mueller-Smith, Daphna Bassok, Julian Betts, Kevin Stange, Celeste Carruthers, Andrew Simon, Tyler Radler, Owen Kay, Jordy Berne, Stephanie Owen, Brenden Timpe, 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 and Brigham Young University. 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

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

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

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

[1]Fortner and Jenkins (2017) suggest positive *and* negative selection because some redshirters are higher achieving and others

outcomes, my contribution suggests that children who wait are negatively selected in third grade achievement. This negative selection in levels is consistent with strategic efforts to start children in school only when they are "kindergarten ready."

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

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

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

are more likely to be in special education. The share in special education is small enough that this analysis would suggest positive selection on average.

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

around kindergarten recommendations is responsible for up to 15% of the income-achievement gap in third grade, validating conjectures that selection around recommendations reinforces learning gaps (Graue and DiPerna, 2000; Deming and Dynarski, 2008) and equity-based discussions in support for requirements (e.g., Illinois, 2019). My contribution is estimating heterogeneity over observed and unobserved dimensions to quantify the equity-efficiency tradeoff that results from strategic selection.

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

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

The remainder of the paper includes the following sections: (2) defining the conceptual and econometric

framework for this research; (3) describing the data and policy context; (4) estimating the effects of waiting with the regression discontinuity and testing for selection in levels and selection on gains between compliers; (5) exploring selection in levels and selection on gains for children who would always wait even when recommended to start kindergarten; (6) presenting suggestive results about mechanisms, the welfare framework, and policy simulations; and (7) containing my conclusion, discussion of results, and possible future research.

2. Conceptual Framework

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

2.1 First Stage: Waiting to Start Kindergarten

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

There are three concrete advantages to defining treatment as waiting to start kindergarten at six. First, waiting is the one behaviorally-relevant decision for parents and policy makers. Although research often studies entry age, relative age, testing age, and their interactions separately, my definition does not separate them because they are not separately manipulable by the agents. Second, this binary treatment avoids well-known identification issues with continuous age-based definitions of treatment (Angrist and Pischke, 2008; Barua and Lang, 2016). Finally, defining treatment as waiting to start kindergarten at six makes it clear that redshirting (waiting when recommended to start), early entry (starting when recommended to wait), and on-time entry are not three categorical "treatments" as previously characterized (Cook and Kang, 2018; Molnar, 2020; Jenkins and Fortner, 2019). Rather they are manifestations of selection into and out of treatment. Appendix C compares my framework with prior work answering both descriptive and causal questions in more detail.

Having defined waiting as the decision of interest, I refer to parents as the main decision makers. While professionals often comment on kindergarten readiness and districts may provide entry guidelines, families

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

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

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

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

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

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

2.1.1 Characterizing Families by Their Reluctance to Wait

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

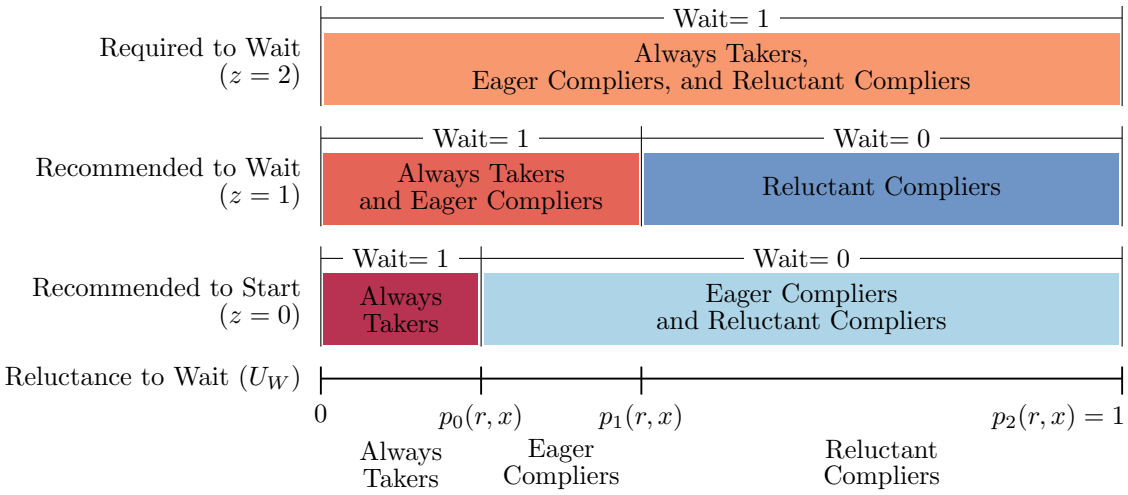
Figure 1 illustrates the connection between these groups and families' innate reluctance to wait, U_W. Imagine a set of families with similar characteristics, x, and birthdays, r. Among them always takers

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

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

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

Figure 1: Reluctance to Wait for Always Takers, Eager Compliers, and Reluctant Compliers



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

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

[7] This is reflected in Figure 1 where never takers are not depicted. I do this because never takers would seem to be violating a requirement, and in the data there fewer than 100 students identified as never takers in each cohort (less than 0.1%)—a number of which may be due to data or measurement error. Dropping these students eases exposition but does not change any results.

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

second, this variation will help identify the nature of selection around kindergarten entry recommendations.

Second, as Figure 1 illustrates, these groups are ordered by U_W which will inform how I compare test scores and the effects of waiting between groups. Specifically, always takers are less reluctant to wait (lower U_W) than eager compliers, who in turn are less reluctant to wait than reluctant compliers. The implications of strategic selection depend on always takers and reluctant compliers—as they *do not* comply with recommendations. The next section leverages the fact that these families are on opposite extremes of U_W to formalize ideas of selection in levels and selection on gains.

2.2 Second Stage: The Effects of Waiting

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

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

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

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

The second function is the "marginal treatment effect" (MTE) function. It defines how the effects of waiting vary between families who are more or less reluctant to wait. I use this function to conceptualize "selection on gains." In the presence of positive selection on gains, families who are more likely to wait

(lower U_W) experience larger test score improvements than those who are less likely to wait (higher U_W), implying τ_{MTE} is decreasing in u. In the presence of negative selection on gains, τ_{MTE} would be increasing in u. In addition to describing the nature of selection around entry recommendations, the information from τ_{MTE} has implications for the equity and efficiency of recommendations and requirements as described in Appendix B.

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

2.2.1 Identification with Birthday Cutoffs

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

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

These means are weighted averages of the underlying MTR functions m_0 and m_1,[9] but unlike the MTRs, many of these means will be identified. To see this I return to the discussion of birthday cutoffs postponed in Section 2.1.

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

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

(LATE) on eager compliers is also identified: $\tau_{ec}(r^*) = \mu_{1,ec}(r^*) - \mu_{0,ec}(r^*)$ as is the selection between eager and reluctant compliers, $\mu_{0,ec}(r^*) - \mu_{0,rc}(r^*)$.

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

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

2.2.2 Extrapolation from Birthday Cutoffs

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

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

To extrapolate, I adopt a parallel trends assumption. I assume that in expectation, children's potential test scores had they waited to enter kindergarten evolve in parallel to their potential test scores had they

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

started kindergarten at age five (for some extrapolation window $[\underline{r}, \bar{r}]$ around a cutoff r^*):

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

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

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

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

My parallel trends assumption implies that x and u are the main drivers of heterogeneity, as do approaches

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

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

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

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

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

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

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

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

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

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

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

3.1 Michigan Public School Data Contain Needed Information

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

In my analysis, I focus on third grade test scores, the nearest-term outcomes to kindergarten entry. Focusing on third-grade scores allows me to investigate selection and effect heterogeneity without worrying as much about differential outcome dynamics between students with different observed and unobserved characteristics. Note that treatment effects on third-grade scores compare students who test at different ages in the same grade (third graders who are nine instead of eight) rather than at about the same age in different grades (nine-year-old students in third grade instead of fourth grade). These within-grade effects are the comparisons of interest for decision makers. For example, students will be tracked into accelerated

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

or remedial paths based on their in-grade scores; report cards and high school transcripts list in-grade performance; states administer high school exit exams in a given grade not at a given age; and students who apply to college will be evaluated against same-grade rather than same-age peers.[17] Fortunately, in Michigan tests are psychometrically calibrated for comparing scores on a given grade's tests across years.

3.2 Michigan Policies Affecting Kindergarten Entry

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

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

3.3 Focusing on the Cutoffs in 2013 Provides the Cleanest Comparisons

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

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

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

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

Second, the two cutoffs identify the effects of waiting on eager and reluctant compliers. Students who turned five after November 1 were *recommended* to wait, but students who turned five after December 1 were *required* to wait. Since students who turned five between November 2 and December 1 were still allowed to start without waiting, the November cutoff identifies the LATE on eager compliers and the December cutoff identifies the LATE on reluctant compliers. Comparing these two LATEs is the first step in characterizing selection on gains.

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

3.4 Descriptive Evidence of Selection

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

First, note that the probabilities of waiting at the November 1 cutoff identify the shares of always takers, eager compliers, and reluctant compliers. Recall that $p_z(r)$, the probability that a children with a given birthday r who faces a recommendation or requirement z, describes the shares of each group. Because these probabilities are identified at the November 1 cutoff, the shares are as well. Panel (a) of Figure 2 illustrates this graphically. For example, among children recommended to start at November 1, I find that $p_0(r_{Nov}) = 0.18$, indicating that the share of always takers at November 1 is 18%. Similarly, among children recommended to wait at November 1, I find that $p_1(r_{Nov}) = 0.40$, indicating that the shares of eager compliers and reluctant compliers are 22% and 60% at November 1.

Given the information about group shares, the reduced form relationship suggests positive selection on gains because the two discontinuities in third-grade math scores are not proportional to the two discontinuities in waiting. As shown in Panel (b) of Figure 2, the probability of waiting jumps by about 0.48 at the December 1 cutoff and only 0.22 at the November 1 cutoff. Absent selection on gains, the effects on eager and reluctant

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

compliers would be equal, and the jump in scores at December 1 would be more than twice the size as the jump at November 1. Nevertheless, Figure 2 shows this is not the case. The jump at December 1 is only about 30% bigger, suggesting that eager compliers (moved by the recommendation at the November cutoff) gained more from waiting than reluctant compliers (moved by the requirement at the December cutoff). In other words, it suggests positive selection on gains.

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

Panel (c) of Figure 2 reports average scores by children's birthdays *and* whether they waited to enter kindergarten and shows jumps in test scores at November 1 even conditional on waiting. Among students who do not wait (dark and light blue lines), scores may increase at November 1, suggesting that eager compliers are negatively selected in relative to reluctant compliers. Among students who do wait (red and orange lines), scores seem to decrease at November 1, suggesting that always takers are positively selected in levels or on gains (or both) relative to eager compliers. At December 1 there is not much of a discontinuity, but this comparison cannot discern whether this is because there is no selection in levels and no selection on gains or because the two net each other out.

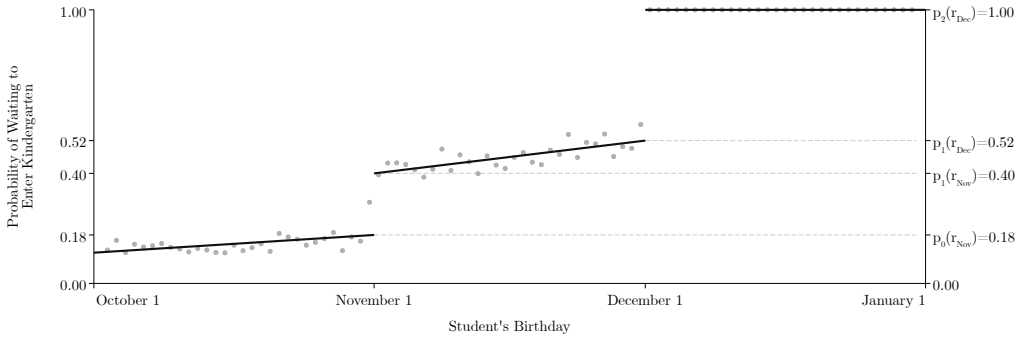
4. RD Estimates of Compliers' Selection in Levels and Selection on Gains

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

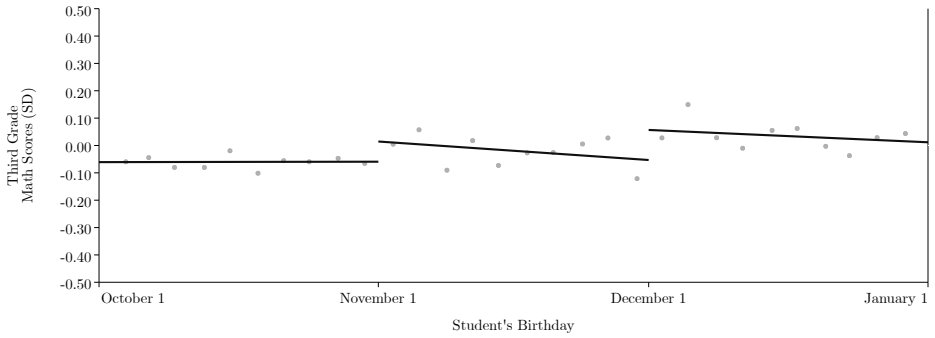
4.1 Both Eager and Reluctant Compliers Benefit from Waiting to Enter Kindergarten

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

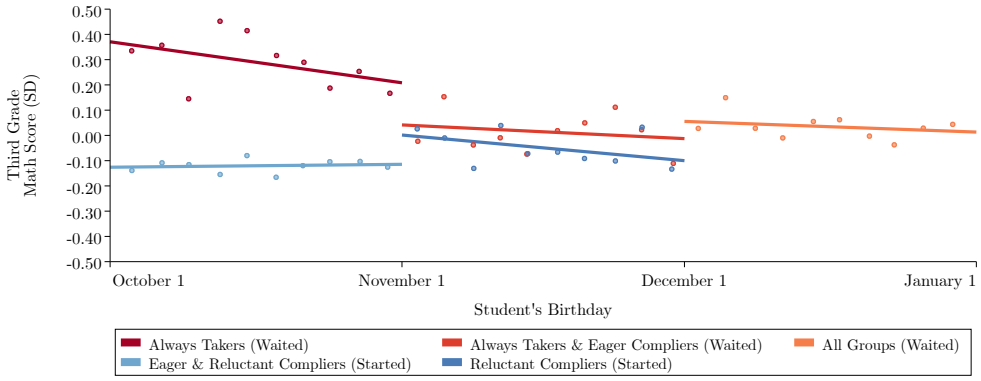
Figure 2: Changes in Waiting and Scores at Cutoffs Suggest Selection is Important



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



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



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

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

instrumental variables regressions at each cutoff (for $k = 1, 2$):

$$Y_i = \tau_k W_i + \beta_k X_{ik} + v_{ik}$$

$$W_i = \gamma_k z_{ik} + \pi_k X_{ik} + u_{ik}$$

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

The main identifying assumption for $\hat{\tau}_1$ and $\hat{\tau}_2$ to be consistent for the target LATEs, τ_{ec} and τ_{rc} (suppressing dependence on r), is that potential outcomes be continuous over both discontinuities. Black et al. (2017) point out that this is an exclusion argument: If z affects outcomes through a mechanism other than W, it would violate the standard LATE exclusion assumption. In my setting this assumption is equivalent to discontinuities in individual potential outcomes at the cutoff. This assumption might be violated if policies or practices led a nonrandom subset of parents to plan births to one side of the cutoffs or encouraged differential childhood investments in children with birthdays to either side of the cutoffs. Appendix E explores these and other potential confounding factors, finding no evidence that this assumption is violated.[22]

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

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

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

Table 1: Waiting to Enter Kindergarten Increases Eager and Reluctant Compliers' Scores

Panel A: November 1 Cutoff	First Stage (RD: Wait to Enter)	Reduced Form (RD: Third Grade Math)	τ_{ec} (Eager Compliers) (Fuzzy RD: Third Grade Math)
Effect	0.225***	0.073*	0.325*
	(0.015)	(0.032)	(0.144)
Slope Before Cutoff	0.002***	0.000	-0.000
	(0.000)	(0.001)	(0.002)
Change in Slope	0.002*	-0.003	-0.003
	(0.001)	(0.002)	(0.002)
Panel B: December 1 Cutoff	**First Stage (RD: Wait to Enter)**	**Reduced Form (RD: Third Grade Math)**	**τ_{rc} (Reluctant Compliers) (Fuzzy RD: Third Grade Math)**
Effect	0.478***	0.111**	0.232**
	(0.014)	(0.034)	(0.071)
Slope Before Cutoff	0.004***	-0.002	-0.003*
	(0.001)	(0.001)	(0.002)
Change in Slope	-0.004***	0.001	0.002
	(0.001)	(0.002)	(0.002)

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

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

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

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

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

I test the null hypothesis that $\Delta_{ec,rc} = 0$ by simultaneously estimating both effects with GMM.[23] Because τ_{ec} and τ_{rc} are estimated at different values of the running variable, the parallel trends assumption is necessary to attribute the difference to selection on gains. Appendix E.2 explain a test of the parallel trends assumption suggesting that parallel trends hold in this setting.

When estimated in the full sample, the difference, $\Delta_{ec,rc} = 0.097$ is economically large, but it is statistically insignificant; however, I find strong evidence for selection on gains from subgroup specific effects. I compare $\hat{\tau}_{ec}(x)$ and $\hat{\tau}_{rc}(x)$ for low-income, high-income, Black, white, female, and male students using the same approach. Table 2 reports the results from these comparisons which include all observations within 90 days of either cutoff to remedy the power limitations of subgroup effects.

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

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

To test for selection in levels I compare the third-grade math scores of eager and reluctant compliers who both start kindergarten at five. As neither group waited, any differences must stem from the fundamental differences between eager compliers and reluctant compliers, recall that I call these differences selection in levels. Whereas estimating selection on gains required an additional assumption to extrapolate over birthdays, selection in levels is identified at the November 1 cutoff with no additional assumptions. I find strong evidence of negative selection in levels on average and among most subgroups but cannot reject the null of no selection among low-income students.

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

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

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

Table 2: Eager Compliers Benefit More than Reluctant Compliers in Some Subgroups

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

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

I test the null hypothesis of no selection in levels $\mathcal{B}_{ec,rc} \equiv \mu_{0,ec} - \mu_{0,rc} = 0$[24] by applying the results from Imbens and Rubin (1997) and Abadie (2002) to identify the expected outcomes of compliers (See Appendix D for details). As such my test is intuitively similar to testing whether never takers have expected outcomes equal to the control complier mean. Equivalent tests for fuzzy RD settings are proposed in both Bertanha and Imbens (2019) and Black et al. (2017).[25]

The estimated selection statistic is large and statistically significant in the full sample and for most subgroups. Table 3 reports the results using a nonparametric block bootstrap by school for inference. In the full sample, $\mathcal{B}_{ec,rc} = -0.42$, and I reject the null hypothesis of no selection at a $p = 0.004$. While the differences are negative and economically meaningful for all groups, there are stark difference in estimate size and statistical significance by racial and socioeconomic subgroups. For example, among low-income students the difference between eager and reluctant compliers' third grade test scores is only 0.12σ ($p = 0.289$), but among higher-income students the difference is over 0.66σ ($p = 0.002$). The differences are larger among

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

[25]Bertanha and Imbens (2019) note that a discontinuity of $\mathbb{E}[Y_i|D_i = 0, r]$ at the RD cutoff violates their strong "external validity" assumption, meaning that the complier estimated effects may not generalize to always takers or never takers. I employ the test to measure selection, a possibility noted in Black et al. (2017) which notes that rejecting the null "constitutes evidence of either selection or violation of the exclusion restriction." This test can also be framed as a special case of a more general test proposed in (Mogstad et al., 2018).

white students compared to black students, but are similar between boys and girls. Substantively, these findings mean that reluctant compliers outperform eager compliers when both groups start at five. In other words, students are negatively selecting in levels into waiting. In addition to showing the economic nature of selection, this finding demonstrates that children who enter early tend to be higher achieving and also suggests that early entrants are much more positively selected than OLS estimates suggest.[26].

Table 3: Untreated Eager Compliers Have Lower Scores than Untreated Reluctant Compliers

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

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

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

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

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

Knowing that there is negative selection in levels and positive selection on gains between complier groups only answers half of the question. Do always takers are select in the same way? This section begins by reinterpreting the RD results in a marginal treatment effects (MTE) framework to show why the nature of selection into redshirting is critical for identifying τ_{at}. Then it and demonstrates that always takers are negatively selected relative to eager compliers, and uses that fact to estimate the effect of waiting to enter on always takers.

5.1 Mapping the RD Results into an MTE Framework

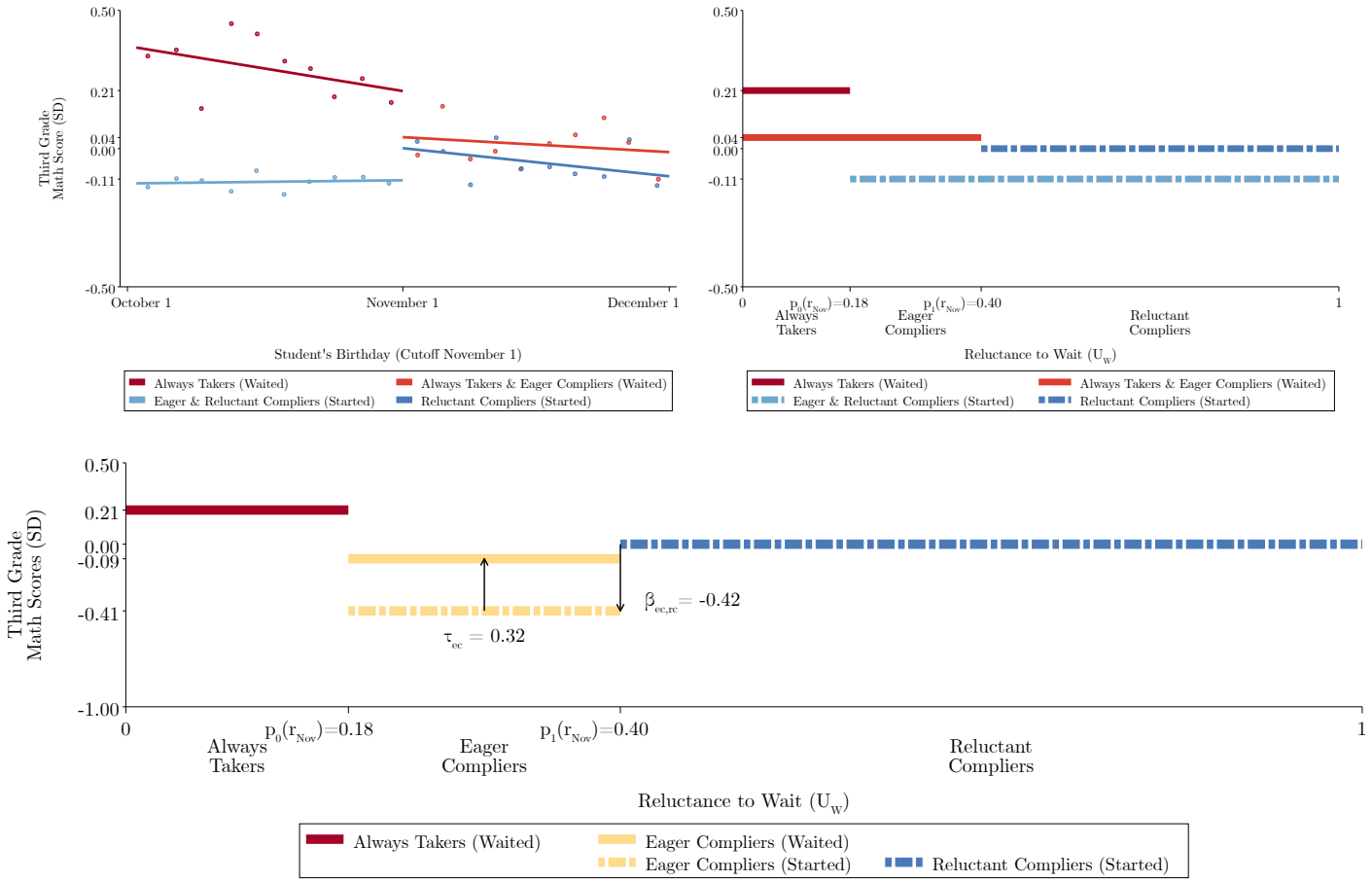
Recasting the results from Section 4 in a MTE framework shows how the scores at the cutoff identify averages of the marginal treatment response and marginal treatment effect curves, $m_0(u, r_{Nov})$, $m_1(u, r_{Nov})$, and $\tau_{MTE}(u, r_{Nov})$. Figure 3 shows these connections graphically in three panels. The top two panels of Figure 3 map the average scores identified by the RD to $\mu_0(u, r_{Nov})$,$\mu_1(u, r_{Nov})$. The top left panel displays average third-grade math scores around the November 1 cutoff separately by whether students waited to enter. The regression lines identify four limits at November 1. The top right panel plots the implied average test scores identified by these limits as the average values of m_0 and m_1 over each group's reluctance to wait, U_W. For example the RD identifies $\mu_{1,at}(r_{Nov}) = 0.21\sigma$. Since 18% of students redshirt at the cutoff, $\mu_{1,at} = \mathbb{E}\left[m_1(u, r_{Nov})|u \in [0.00, 0.18]\right]$, so I plot a line at 0.21 over $U_W \in [0.00.018]$. The other three line segments report the averages for the other groups: always takers and eager compliers who wait, eager and reluctant compliers who start, and reluctant compliers who start.[27] Each is an average of the relevant marginal outcome function $\mu_1(u, r_{Nov})$ or $\mu_0(u, r_{Nov})$.

This interpretation of the results also identifies portions of the MTE curve, $\tau_{MTE}(u)$ and MTR curve measuring selection in levels, $m_0(u, r_{Nov1})$. This is reflected in the bottom panel of Figure 3 which plots the average outcomes over U_W after recovering the complier means. This presentation of the results makes τ_{ec}, τ_{rc}, and $\mathcal{B}_{ec,rc}$ visible. τ_{ec} reflects the average values of τ_{MTE} over $U_W \in [0.18, 0.40]$ and $\mathcal{B}_{ec,rc}$ the difference in test scores after starting kindergarten at five between $U_W \in [0.18, 0.40]$ and $U_W \in [0.40, 1.00]$.

Mapping the RD results into an MTE Framework also makes it visually apparent that identifying the causal effect on groups other than eager compliers requires additional information or assumptions. I need more information because the November cutoff has no information about the scores of reluctant compliers if they wait to start kindergarten or about always takers if they start at five. Fortunately, the December cutoff gives me information about the scores of reluctant compliers who wait, but I will need to make an

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

Figure 3: Reduced From Shows Effects, Selection, and Heterogeneity



Note: This figure displays average standardized third-grade math scores over two different dimensions to illustrate the mapping between the two. In both figures the sample is made up of 15,066 students who entered kindergarten in a Michigan public school, who have birthdays within thirty days of November 1 in 2013, and for whom I observe third-grade math scores. The panel on the top left is an RD plot of outcomes separated by treatment status, similar to Panel (c) in Figure 2. Instead of showing average outcomes for all students at a given index value, this graph plots those average outcomes separately for treated students (who waited to enter kindergarten) and untreated students (who did not wait). Average test scores are reported by three-day bins for students who did and did not wait, and regression lines are displayed for each subgroup. Lines of best fit are also displayed to visualize how the limits of $\mathbb{E}[Y_i|d, z_1]$ are estimated. The panel on the top right maps the expected performance of each group as identified by those limits at November 1 onto the support of their unobserved cost U_W. In this graph the average outcomes of treated groups are displayed in solid lines, and the average outcomes untreated groups are displayed with dashed lines. The bottom panel illustrates the local average treatment effect (LATE) and selection between eager and reluctant compliers implied by these means at the November 1 cutoff.

assumption about the test score of always takers had they entered without waiting. Although it is common to make *a priori* assumptions about the untreated outcomes of always takers, the following section empirically tests the nature of selection in levels instead. Identifying how always takers select in levels will inform what assumptions are reasonable to identify the treatment effects on always takers and whether or not they also positively select on gains.

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

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

5.2.1 Always Takers Receive More Special Education Services and Testing Accommodations

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

The results show large differences between groups and produce perplexing results under no selection or positive selection in levels. Table 4 reports the average outcomes of always takers, compliers who wait, and compliers who enter without waiting. Despite having the highest third-grade math scores (see Figure 3), always takers have higher rates of testing accommodation, non-testing,[29] and special education service receipt.[30] These results are also consistent across demographic subgroups (see Appendix Table F.8) and suggest negative selection in levels because generally it is low achieving students who receive these accommodations.

In addition to the fact that always takes have worse level outcomes than eager compliers, only large

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

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

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

Table 4: Early Elementary School Outcomes Suggest that Always Takers Are Negatively Selected

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

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

and unintuitively signed treatment effects on always takers could rationalize positive selection in levels in these outcomes. Positive selection in levels would suggest that always takers who start kindergarten at age five should experience lower rates of accommodation, non-testing, and special education than compliers who start at age five. But given the observed outcomes for always takers who wait would then imply that waiting to enter kindergarten increases the prevalence of these outcomes in always takers by about 100%; whereas for compliers, waiting to enter kindergarten *reduces* the likelihood of missing a test and of receiving special education by around 20%.[31] These differences in sign are statistically significant and constitute strong evidence that always takers are *not* positively selected in levels relative to compliers. Alternatively, under negative selection in levels, these are exactly the relationships we would expect to see.

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

[31] Finding that waiting reduces special education assignment is not unique to my sample. See also Elder (2010); Evans et al. (2010); Layton et al. (2018); Dee and Sievertsen (2018); Sharpiro (2020) for other examples.

large enough to suggest positive selection in levels on average, the following subsection provides additional evidence of negative selection in levels from third-grade achievement.

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

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

The comparison of interest exploits variation in average achievement as birthdays approach the cutoff. In a sharp RD design where all units to the left of the cutoff are untreated, any change in average untreated outcomes as the running variable r increases captures the direct effect r (or if r and x are not independent any changes conditional on x). This intuition changes in a fuzzy RD because as r increases, the share of always takers increases—the upward sloping lines in the first stage (Panel (a) Figure 2). In a fuzzy RD, the changing share of always takers implies that changes in average untreated outcomes to the left of the cutoff reflect both the direct effect of r as well as the changing composition of the still untreated group.

Because changing scores reflect the direct effect of r as well as the changing composition, if a researcher knows something about the direct effect, the change in scores can reflect the selection in levels revealed by the compositional change. For example, in the case of elementary school achievement, it is well-known that students with later birthdays score worse on tests (conditional on waiting). This is called an "age at test" effect and is visible in the reduced form (Panel (b) of Figure 2). If we observed the scores of children who started kindergarten at five increasing instead of decreasing in some range of r it must be due to compositional changes and would reveal negative selection in levels for the range of $p_0(r)$ where test scores were increasing. Interestingly, this is exactly what Figure 3 revealed.

Figure 3 showed that average test scores of students who start are increasing, implying that always takers are negatively selected. Between October 1 and November 1, there is a large increase in the probability of waiting. Inducing lower-achieving students to wait is increasing the average scores in the remaining group who starts at five. A linear regression of test scores on this window (with controls for lower-income, black, and female) reports a positive slope and rejects a two sided test of the slope being zero at $p = 0.046$ level. Because this result is somewhat sensitive to bandwidth and covariate inclusion, I explore an alternative test in Appendix D based on changes in slope of test scores over birthdates that yields the same results and demonstrates that always takers are negatively selected in levels on average and among higher-income, white, and female subgroups.

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

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

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

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

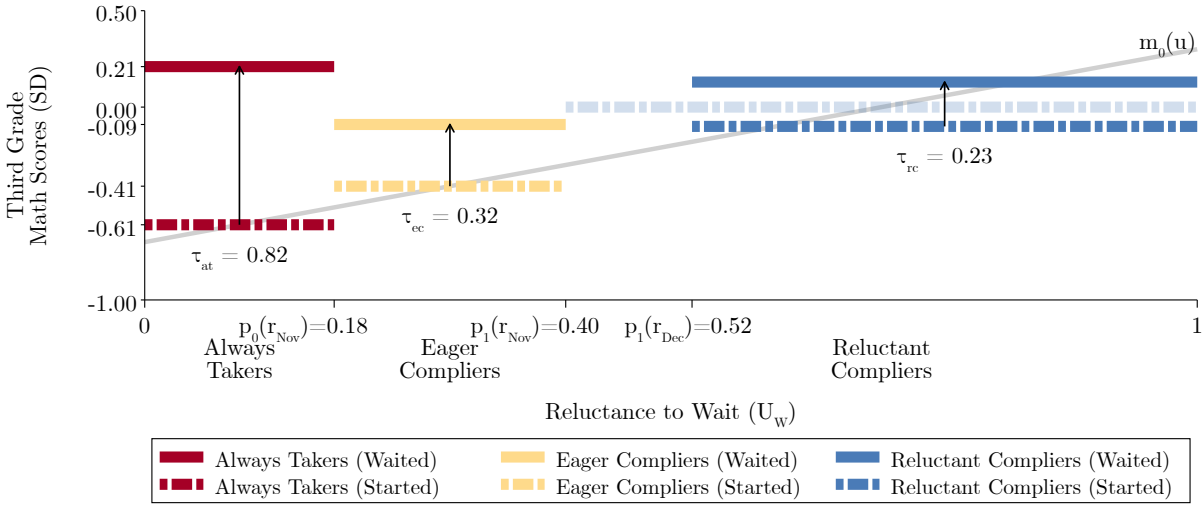
$$m_0(u, r, x) \equiv \alpha_0(x) + \beta_0(x)u_i(x) + v_i \text{ with } E[v_i|u, x] = 0$$

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

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

I find that always takers at the November 1 cutoff benefit enormously from waiting to enter kindergarten, suggesting more positive selection on gains. Figure 4 combines earlier results with the effects on always takers to illustrate the selection on gains. This figure adds the scores of reluctant compliers at December 1 over their values of U_W. Then by plotting the showing the linear $m_0(u)$ over the support of U_W and the implied average scores of always takers had they not waited ($\mu_{0,at} = -0.63$), Figure 4 visualizes the gains to always takers, $\hat{\tau}_{at} = 0.84\sigma$. This large effect is significantly different from zero and from τ_{ec} at a $p = 0.001$ level (Table 5 has all standard errors and p-values as well as results from a bounding assumption presented in detail in Appendix E. Taken together it is visually apparent that the three shrink grow monotonically as the reluctance to wait increases. In other words it clearly displays positive selection on gains.

Figure 4: Late Entrants Benefit Enormously from Waiting to Entering Kindergarten



Note: This figure graphically illustrates the average treatment effect for always takers at the cutoff τ_{at}, which is recoverable from the limits of student achievement at the cutoff and the ancillary assumption of linearity in untreated outcomes. The sample is made up of 15,066 students who entered kindergarten in a Michigan public school, who have birthdays within thirty days of November 1 in 2013, and for whom I observe third-grade math scores. Average outcomes for treated and untreated compliers are backed out of observed data and choice probabilities. For block-bootstrapped standard errors see Table 5.

In addition to showing that always takers benefit from waiting on average, I explore heterogeneity across different demographic groups by splitting the sample on observable characteristics x. A key advantage to nonparametrically estimating heterogeneity by observables is that it allows for different patterns of selection

on gains across groups. As Cornelissen et al. (2016) point out, splitting the sample by x and estimating effects separately is the ideal way to estimate marginal treatment effects when there is strong enough support. Splitting the sample also removes the need to make common assumptions of additive separability between x and u, as these assumptions imply that all groups must select uniformly on gains.

After splitting the sample, I find evidence that only certain types of always takers are positively selected on gains. Table 5 shows that not all groups positively select on gains. For example, the treatment effect on always takers is larger than the effect on eager compliers (which in turn is larger than the effect on reluctant compliers) among the white children and children from higher-income families, but there are not significant differences for other children. There is strong evidence of selection on gains among girls and boys. Table 5 also reports the full results p-values of tests for heterogeneity under weaker assumptions of monotonicity rather than linearity in untreated outcomes with somewhat larger p-values but qualitatively similar results.

Table 5: Only Some Demographic Groups Are Positively Selecting on Gains

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

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

These results would be consistent with the explanation that narrative that regardless of demographics, parents tend to be aware of which children are at risk of underperforming in kindergarten. That said, the varying slopes of the selection in levels may suggest that certain parents respond more strongly to the possibility of poor performance: high-income parents seem to be especially responsive. Simultaneously, the lack of selection on gains suggests that identifying children who are at risk for underperforming is not

sufficient to enable them to succeed, they need to actually benefit from those decisions to wait. I explore possible mechanisms and explanations for this in the subsection that follows.

6. Discussion, Extensions, and Policy Implications

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

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

Before turning to equity and efficiency, this subsection explores the mechanisms through which the observed patterns of selection may operate. The absence of selection on gains I find is not the typical reverse-Roy sorting where low-income parents fail to invest in the children who would benefit the most;[33] rather, the results in Table 5 suggest that waiting does not make as much of a difference for the low-income students. In other words there are not large gains to sort on. This section formally explores these differences and possible drivers. I find that low-income students benefit much less on average from waiting a year to enter kindergarten and show suggestive evidence that this is the result of heterogeneity in the preschool investments made in the intervening year (as also suggested by the tapering of gains over time, see Elder and Lubotsky, 2009).

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

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

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

Note that estimating τ_{ATE} this way requires extrapolating from r_{Dec} to r_{Nov}. This extrapolation is

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

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

I find that the ATE is smaller for low-income students than for higher-income students. The first column of Table 6 reports the results. The average effect on higher-income children in 0.48σ, whereas it is only 0.16σ for the low-income children. This difference is significant at the $p = 0.05$ level and means that on average higher-income children benefit three times more from waiting to enter kindergarten than do low-income students. The difference between the average treatment effect on male and female students is inconsequential, and the difference between black and white students is large but falls just shy of statistical significance at a $p = 0.10$ level.

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

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

variation in the decision to wait to enter kindergarten, not in participation in the GSRP.

Analyzing the GSRP yields suggestive evidence that the differences in ATEs across groups are mainly driven by different quality investments in the year before kindergarten. The main evidence for this claim is the fact that while the ATE for low- and higher-income students are very different in the full sample, the effects are almost identical on those who participate in GSRP. The second and third columns of Table 6 detail the results. For comparisons by economic status and by race, the ATE is statistically indistinguishable among students who did GSRP the year before kindergarten; in fact, the difference between the higher-income ATE and the low-income ATE is less than 0.02σ. On the other hand, among children who do not participate in GSRP, the differences in the effects of waiting are extreme: the large for higher-income children and white children dwarf the statistically insignificant effects on low-income and black children (both differences are statistically significant).

Table 6: Average Treatment Effects Are Larger for Higher-Income Families, Except Among Children in Public Pre-K

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

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

These results are suggestive of sharp differences in the quality of the human capital investment entailed in waiting a year to enter kindergarten. Because preschool enrollment decisions are not random, these results must be interpreted as differences in the average effects for students who choose to (or not to) participate in GSRP, and as low-income children are more likely to qualify, selection is likely not uniform across groups.

The patterns suggest that differences in the effects of waiting are not driven dynamic complementarities. In the presence of such complementarities we would expect the higher-income children to benefit more from a given preschool program because they have had more intensive investments earlier on.

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

6.2 Measuring the Efficiency and Equity Implications of Strategic Selection

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

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

The first two columns of Table 7 show how banning strategic selection would lower scores but shrink

Table 7: Allowing Strategic Selection Raises Scores But Widens Gaps

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

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

gaps. To estimate these effects I impose two assumptions: I assuming that the effects on reluctant compliers estimated at December 1 reflect the effects on all reluctant compliers (as when calculating the ATE) and that the effects on always takers estimated at November 1 reflect the effects on all always takers. Note that if marginal treatment effects are monotonic over U_W, these assumptions minimize the efficiency gains of strategic selection (relative to a policy with requirements). Currently the income-achievement gap is about 0.79 standard deviations. Banning redshirting and early entry reduces this gap by about 18%. The gap is closed in both directions. Scores among children from lower-income families increase because the gains to reluctant compliers more than compensate the losses to always takers (since there are many more reluctant compliers than always takers). But scores among children from higher-incomer families are lowered tremendously because the gains to reluctant compliers are small, and the losses to always takers are large. Also note that in addition to lowering average scores, the allowing strategic selection is revealed preferred, so the requirement policy imposes large costs on parents. These results highlight a real equity-efficiency tradeoff between requirements and recommendations.

A policy simulation using the LATE from eager compliers would get this efficiency equity implications completely backwards. The third column of results from Table 7 shows this result by assuming that the eager-complier LATE for higher- and lower-income families is the effect on all students. Because there is no selection on gains (positive or negative) among lower-income families, extrapolating using the LATE to always taker and reluctant compliers does not change much among that group. On the other hand, among higher-income families the eager-complier LATE largely overstates the benefits of a requirement to reluctant compliers and understates the losses to always takers. In net this means that using the LATE for policy analysis gives the wrong answer. It says that a requirement will raise scores for both groups, but because

the gains are larger to higher-income families, it will widen gaps. This answer is wrong on both accounts and shows the importance of allowing for selection on gains especially for questions about allowing strategic noncompliance with treatment recommendations.

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

7. Conclusion

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

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

The fact that children who are negatively selected are more likely to wait to enter kindergarten does not dismiss the equity concerns about strategic selection. In fact, my results show that these concerns are well founded: Despite being negatively selected on average, after waiting, always takers (including academic

redshirters) are among the highest performing students. This pattern of selection on gains is concentrated among higher-income families, implying that in the status-quo allowing selection perpetuates racial- and income-based gaps in achievement.

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

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

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

While this paper focuses on the economic nature of selection into waiting, it is entirely focused on short term outcomes. Applying my framework to research on longer-term outcomes would be important and policy relevant. I show that reluctant compliers are positively selected in levels compared to eager compliers, and that eager compliers are positively selected relative to always takers—but this finding is limited to third-

grade test scores. In the same way that treatment effects on compliers have been shown to fade out over more extensive time horizons, the magnitude (or even direction) of selection need not be constant across outcomes measured at different lengths of time since treatment. This too seems like an important avenue to consider the effects of redshirting in the long run.

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

References

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

Adhvaryu, A., S. Bednar, T. Molina, Q. Nguyen, and A. Nyshadham (2020): "When It Rains It Pours: The Long-Run Economic Impacts of Salt Iodization in the United States," Review of Economics and Statistics, 102, 395–407.

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

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

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

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

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

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

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

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

Black, D., J. Joo, R. LaLonde, J. A. Smith, and E. Taylor (2017): "Simple tests for selection: Learning more from instrumental variables," Tech. rep.

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.

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.

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," Tech. rep., National Bureau of Economic Research.

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.

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.

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.

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.

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.

Fryer, R. G. (2017): "The production of human capital in developed countries: Evidence from 196 randomized field experiments," in Handbook of economic field experiments, Elsevier, vol. 2, 95–322.

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.

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation 1," Econometrica, 73, 669–738.

Hemelt, S. W. and R. B. Rosen (2016): "School entry, compulsory schooling, and human capital accumulation: evidence from Michigan," The BE Journal of Economic Analysis & Policy, 16.

Illinois, G. A. (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.

Ito, K., T. Ida, and M. Tanaka (2021): "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice," Tech. rep., National Bureau of Economic Research.

Jenkins, J. M. and C. K. Fortner (2019): "Forced to Redshirt: Quasi-Experimental Impacts of Delayed Kindergarten Entry," Tech. rep., Ed Working Papers.

Jimerson, S. R. (2001): "Meta-analysis of grade retention research: Implications for practice in the 21st century," School Psychology Review, 30, 420–437.

Johnson, R. C. and C. K. Jackson (2019): "Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending," American Economic Journal: Economic Policy, 11, 310–49.

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. (2022a): "Behavior within a Clinical Trial and Implications for Mammography Guidelines," Tech. rep., Review of Economic Studies, forthcoming.

——— (2022b): "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," Tech. rep., Review of Economics and Statistics, forthcoming.

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

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," The American Economic Review, 604–620.

Layton, T. J., M. L. Barnett, T. R. Hicks, and A. B. Jena (2018): "Attention Deficit–Hyperactivity Disorder and Month of School Enrollment," New England Journal of Medicine, 379, 2122–2130.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," Journal of Economic Literature, 48, 281–355.

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.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," Econometrica, 86, 1589–1619.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2020): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," Tech. rep., National Bureau of Economic Research.

Molnar, T. L. (2020): "The Impact of Academic Redshirting on Student Outcomes and Mental Health: Evidence from Hungary," Tech. rep., https://www.dropbox.com/s/bdk9uytrljrbne0/Molnar-Timea-Laura-Redshirting.pdf?dl=0.

Morrison, F. J., D. M. Alberts, and E. M. Griffith (1997): "Nature–nurture in the classroom: Entrance age, school readiness, and learning in children." Developmental Psychology, 33, 254.

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.

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," Tech. rep., SSRN.

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

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, 5, 2332858419848442.

Sharpiro, A. (2020): "Age at time of kindergarten entry and special education service receipt," Working paper.

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

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

A. Data Appendix

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

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

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

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

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

B. Implications of Strategic Selection for Efficiency and Equity

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

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

that because one allocation is more efficient than another if parents prefer it *and* it results in higher scores, there may be some ambiguous comparisons (for example an allocation that is revealed preferred but with low average scores versus one that is not preferred but has higher scores), so allocations are only partially ordered.

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

A second restriction imposed by this model of welfare is its focus the social benefits on average test scores. This focus has two main implications. First, it assumes that test scores capture the social benefits not present in family utility. This assumption rules out examples where a child grows up and regrets the forgone year of earnings despite being higher achieving in school or where she experiences life-long returns from noncognitive gains from waiting beyond what was captured in test scores.[34] A second, more subtle implication is that this assumption limits the extent to which waiting generates spillovers in test scores. It does not rule out the possibility of spillovers (in fact classrooms do benefit from peers who wait, see Peña, 2017); however, it limits the roll of spillovers to operate only through the channel of average achievement not other outcomes or higher-order moments. In most cases this is reasonable, although there is evidence of small increases in crime among students who wait which are not captured by test score gains (Cook and Kang, 2016).

Because social planners and policy makers may care about the outcomes of different groups, a utilitarian concept of efficiency is likely insufficient for discussing welfare. To address this concern I also introduce a simple criterion for equity. Equity for students of type $X = x$ is measured in average differences in realized test scores $\mathbb{E}[Y|X = x] - \mathbb{E}[Y|X \neq x]$, and allocations can be compared by measuring the resulting change in achievement gaps. As such, an allocation is more equitable for students of type x if this gap is smaller. Changes in the size of the gaps arise because different allocations change who waits and because different types of children x have different baseline likelihoods of waiting and different expected gains from doing

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

so. Note that whereas the definition of efficiency included both indirect utility and student achievement, the concept of equity is focused only on achievement. This is consistent with how inequities are defined in income, tax incidence, labor market outcomes, and other settings.

These ideas of efficiency and equity feature in policy conversations surrounding kindergarten entry recommendations and requirements—although rarely using those words. For example, the argument that there should be requirements that force children who turn five after the cutoff has both an efficiency and and equity argument. Strategic selection on this margin lowers average scores, and if the children who would benefit the most from "the gift of time" are less likely to wait, these losses could be large.[35] Furthermore, if these families come from systematically disadvantaged groups, forgoing gains by choosing to start when recommended to wait would widen inequities. These rationales may be behind the policies in 15 US states and many school districts that have strict requirements to wait for children with birthdays after the cutoff.

On the other hand, conversations about academic redshirting often feature both equity and efficiency arguments. When discussing requirements that would force children who turn five before the cutoff to start and not wait (denying treatment to would-be always takers), it is often pointed out that this type of selection is driven by highly educated and often high income families. The assertion is that allowing these families to select around the recommendation to start exacerbates persistent gaps is an equity-based justification for an entry requirement rather than a recommendation. On the other hand, another common assertion is that because these children tend to come from privileged backgrounds, they would do well no matter what. This is an argument about selection that implies that there are not test-score gains from allowing the strategic selection around the recommendation which is related to the ideas of efficiency and average scores.[36]

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

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

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

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

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

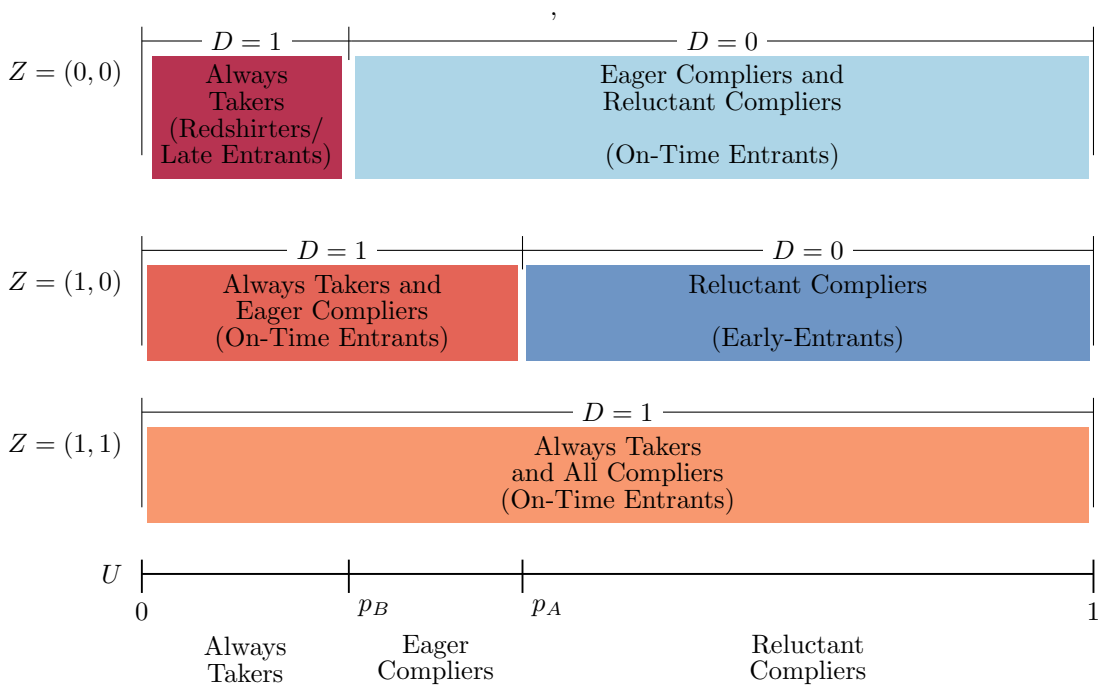
The key insight behind my conceptualization of treatment is that each student faces the choice between two entry years. A student with a given birth date can either enter this year (usually at or just below five years old) or wait to enter kindergarten next year (usually at or just below six). I define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

This definition of treatment and the resulting potential outcomes work better than using the traditional entry groups (like those used in Bassok and Reardon, 2013; Fortner and Jenkins, 2017, for example). Using "early" and "late" entry as treatments or potential outcomes relative to "on-time" entry is problematic because the groups are dependent on assignment. Consider a student who turned five on November 1 and is choosing between entering in 2013, as assigned, or waiting to enter in 2014 (redshirting). There is no material difference between her choice and the choice of a student who turned five on November 2 and who was assigned to wait. If they wait, both students would enter in the same year at the same absolute and relative age. This comparison makes it clear that redshirting (waiting when assigned to start) is not a distinct treatment from complying with an assignment to wait; rather the difference is in the selection into waiting.[37] Figure C.1 illustrates the inconvenient implications of the early, on-time, late framework. None of the groups are mutually exclusive. In fact on-time entrants are a combination of always takers, eager compliers who wait, eager compliers who enter, reluctant compliers who wait, and reluctant compliers who enter. Furthermore, not all always takers are late entrants nor are all reluctant compliers early entrants because of randomization.

Defining treatment as waiting a year to enter kindergarten also serves as a better treatment variable than using continuous "entry age." There are intuitive and econometric reasons for this. The intuitive reason is that waiting or not waiting is the choice that decision-makers face. Knowing the causal effect (while interesting, academically perhaps) is not behaviorally relevant because parents and policy makers cannot

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

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



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

manipulate it separately from testing age, relative age, and their interactions. For them the effect of interest is not that of being a year older (separate from everything that comes with it), but of waiting (along with everything that comes with it).[38]

Econometrically, my definition of treatment overcomes two problems with using "entry age." The first is an issue of fundamental unidentifiablilty, reflected from the intuition. Just as decision-makers generally can not separately manipulate entry age, econometiricians generally cannot separately identify the effects of entry age from testing age (Angrist and Pischke, 2008).[39] The second econometric issue is a failure of monotonicity. When there is noncompliance, using continuous age with either a month (or day) of birth instrument or the variation around the birthday cutoff violates the monotonicity assumption necessary to identify the school-entry LATE (Barua and Lang, 2016). A binary treatment variable like waiting to enter kindergarten overcomes these issues, something Black et al. (2011); Dhuey et al. (2019) apply in their respective settings.

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

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

D. Econometric Appendix

D.1 Recovering Control Complier Means

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

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

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

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

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

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

D.2 Functional Form Test for Negative Selection of Always Takers

My third approach to test for selection in levels exploits variation in average achievement of students who enter without waiting as the share of redshirters increases. As the birthdays approach to November 1 (from the left), the share of redshirters increases monotonically as "marginal" always takers are induced to redshirt.[40] As the probability of waiting increases, the composition of students who enter without waiting changes, and so the slope of average scores among children who enter ($\mathbb{E}[Y_i|D = 0, r]$) captures both the

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

causal effect of being one day older on scores and the changing composition. This has implications for selection in levels. If the "marginal" always takers are positively selected in levels (relative to the students who enter), inducing them to wait will reduce the average scores among the remaining students who enter. In this scenario, the composition changes more as r increases, implying that the average scores should fall more quickly than the causal effect of r. On the other hand, if marginal always takers are negatively selected in levels, the compositional change will increase the average, implying a slope greater than the causal effect of r. The challenge is that the true effect of r and the compositional change are not separately identified in general.

Child development offers a theoretical insight that I turn into a test for negative selection. Because students enter kindergarten at young ages, the effect of an additional month of "entry age" is not thought to be uniform. In fact, the theory suggests that the additional time should be more valuable to younger students than to older students: A month of maturity and experience is relatively more to a five year old than to a six year old (Deming and Dynarski, 2008).[41] In my context theory implies two things. First, the slope of $\mathbb{E}[Y_i(0)|r]$ (the causal effect) should always be negative—students with later birthdays enter younger and perform worse—and second, the slope of $\mathbb{E}[Y_i(0)|r]$ should be more negative closer to the cutoff—younger students would benefit more from a given change in absolute age.

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

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

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

I find that "marginal" always takers are negatively selected relative to compliers. Table D.1 shows the results. Both b_2 and b_3 are negative, as suggested by the theory, but the effect in March-July (-0.0008)[42]

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

[42] Because there is so little noncompliance in the March-July region, the slope of $\mathbb{E}[Y_i|D=0, R=r]$ should be a relatively good estimate of $\mathbb{E}[Y_i(0)|R=r]$ in that region. Interestingly, the other region with very little noncompliance (December-February) has a similar slope for treated outcomes (-0.0007). This is what the developmental theory would suggest, as a relative age effect would be similar for a March student who enters and a February student who waits. The fact that it is largely stable on the

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

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

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

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

E. Empirical Robustness Checks

E.1 Regression Discontinuity Robustness Checks

This section explores robustness of the regression discontinuity results to potential pitfalls. Table E.2 shows that the first stage, reduced form, and fuzzy RD relationships are not sensitive to the regression specification. In both panels column (1) reports the main specification from the paper. Columns (2) and (3) change the bandwidth. Shrinking the bandwidth reduces power, but gives similar (possibly larger) results. Widening

December-July window also suggests that the parallel trends assumption employed in Section 4 to compare the November and December LATEs is quite reasonable. In fact, it may well be true along the whole support.

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

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

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

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

the bandwidth (to the relevant side) increases precision slightly and gives similar (possibly smaller) results. Columns (4) and (5) show that the uniform kernel gives similar results as a triangular and epanichokov kernels. Colums (6) through (8) explore other polynomial approximations of the conditional expectation of scores over birthdays. Using levels shrinks the estimates considerably, but only if done without shrinking the bandwidth. Using a quadratic term decreases power but suggests fairly similar (possibly larger) effects.

Since all of the regressions in the paper are performed without controls, Table E.3 shows that the first stage, reduced form, and fuzzy RD relationships are not sensitive to the regression specification. The table focuses on the low-income high-income results to demonstrate that the selection on gains looks similar persists when controls are added. Controls included race, sex, poverty, English language learner, enrollment by school of choice, neighborhood characteristics (percent black, percent Hispanic/other-nonwhite, employment rate, median household income, percent with no high school degree, percent with a bachelors or more), and an interaction of low-income and sex. Including these controls did not do much to the effects, maybe shrinking them toward zero. Including (kindergarten) school fixed effects may reduce the eager complier LATE a little and possibly increases the reluctant complier LATE, but the differences are imprecise.

A key assumption in the regression discontinuity framework is that potential outcomes are continuously distributed around the cutoffs. One violation of this assumption could be if different types of students tend

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

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

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

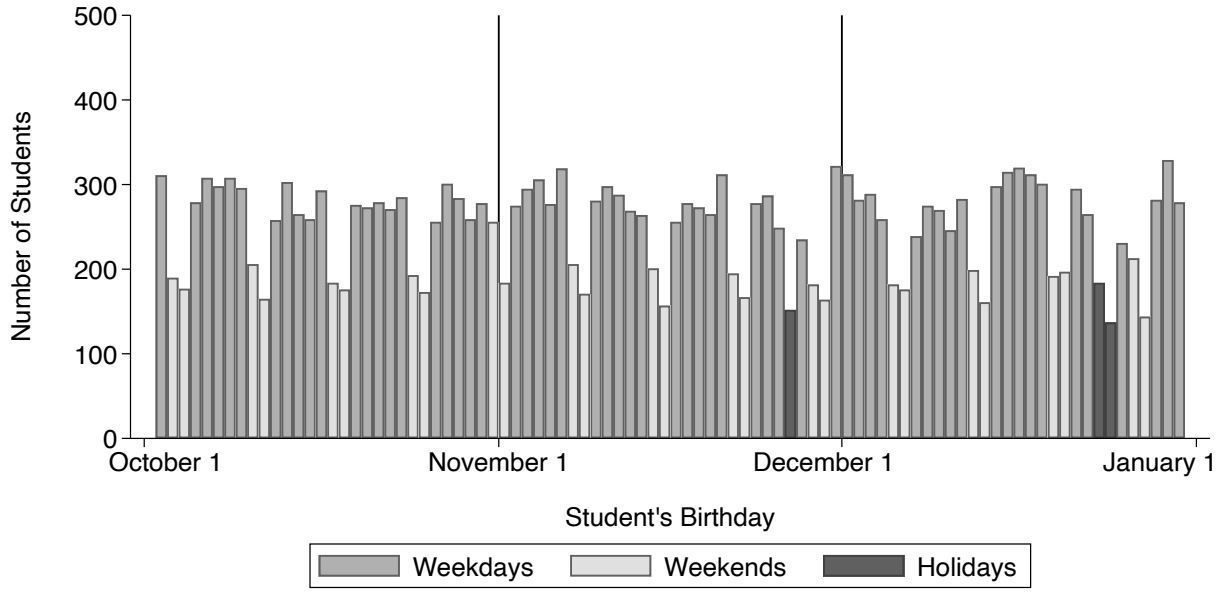
Note: This table compares estimates of the average effect of waiting using different RD specifications. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within 90 days of either cutoff;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

to have birthdays on either side of the cutoff. Because assignment is not enforced it seems unlikely that a significant number of parents strategically plan births to fall on one side of the cutoff or another; however, other factors may influence birth timing. To explore this Figure E.2 reports the counts of students with each birthday.

It is immediately apparent from Figure E.2 that birthdays are not uniformly distributed. For example, children are less likely to be born on weekends and holidays. If these patterns are unrelated to baseline family characteristics, this non-uniformity would not be a problem; however, it is well known that more affluent families are more likely to have their children on weekdays (mostly because of schedule inductions). In my main analytical sample students born on Saturdays are 3.9 percentage points more likely to be low-income; students born on Sundays are 5.6 percentage points more likely to be low-income; and students born on Thanksgiving, Christmas Eve, or Christmas are 6.9 percentage points more likely to be low-income—all relative to students born on non-holiday Mondays (0.54). This is concerning since neither cutoff falls in the middle of the week. As a result the limit could be biased towards an unrepresentative sample. Fortunately, Table E.3 already suggested this was not a problem in practice, and Table E.4 shows that the results are not sensitive to the exclusion of students born on weekends.

A second way to explore the assumption of continuity at the cutoffs is to explore covariates. Figure E.4 displays control variables plotted over the support of birthdays. Overall these comparisons suggest that students born right before and after each cutoff are very similar. When run as regressions only two of the twenty four comparisons are significantly different at the $p = 0.05$ level: low-income at the December cutoff

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



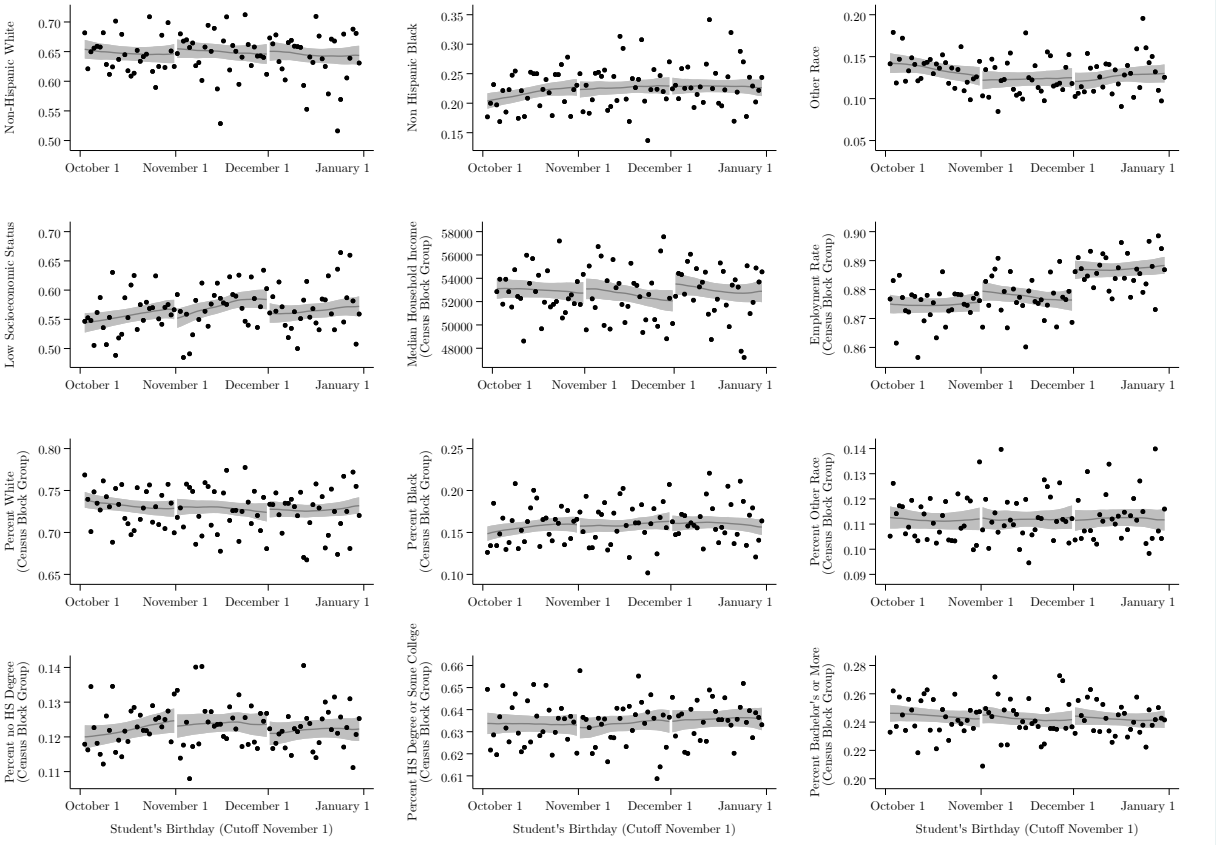
Note: This figure reports the count of students with each birthday. It also shows whether the student was born on a weekend or national holiday. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning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

and and census block group employment rate at the December cutoff. Both of these differences are visible in Figure E.4. I do not know why children who turn five right after December 1 have 1.1 percentage point higher employment rates in their census block groups, but the difference is not present or significant in the low-income or higher-income subsamples, suggesting that conditional on the split there are no differences and that in the whole sample if any differences are present, they are small.

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

The first threat from from Michigan's Great Start Readiness Program (GSRP). The state of Michigan operates GSRP as a program for four-year-old children with special needs, who come from lower-income or divorced families, or who meet other risk criteria. The threat to identification is that for my main analysis cohort students were eligible for GSRP in 2012 if they turned five between November 2, 2012 and December 1, 2013 (older children could have participated in 2011). The threat to identification is that if GSRP takeup is discontinuous at November 1 *and* GSRP affects third grade test scores, the potential outcomes would not be comparable on either side of the cutoff. Indeed, participation in GSRP increases by about five percentage

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



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

points at the November 1 cutoff (among low- and higher-income subgroups), but it does not jump at the December 1 cutoff as would be expected. Table E.4 shows that dropping GSRP participants does not erase the treatment effect in the full sample. Interestingly, conditioning on GSRP and income does not change the selection on gains among high-income or the lack thereof among low-income (if anything low-income GSRP participants select negatively on gains into waiting).

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

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

Unfortunately, there is no comprehensive record of which schools and which districts had requirements like this. I do however know which schools offer official developmental kindergarten programs starting in 2014. This is a superset of the set of schools that force early entrants to take developmental kindergarten. Figure E.4 shows that restricting the sample to schools that do not have these programs is sufficient to drive the discontinuity in taking kindergarten twice to zero (which could be over controlling if eager and reluctant compliers do have different propensities to take kindergarten twice). Table E.4 reports the regression results from dropping these schools. Both the eager complier and the reluctant complier LATEs shrink a little in this specification. Since about 20% of the students in my sample attend these schools, the estimates are also a good deal less precise. The subgroup analyses are not reported, but show the same trends as in all the other specifications.

Taken together these results suggest that both sets of Local Average Treatment Effects are well identified

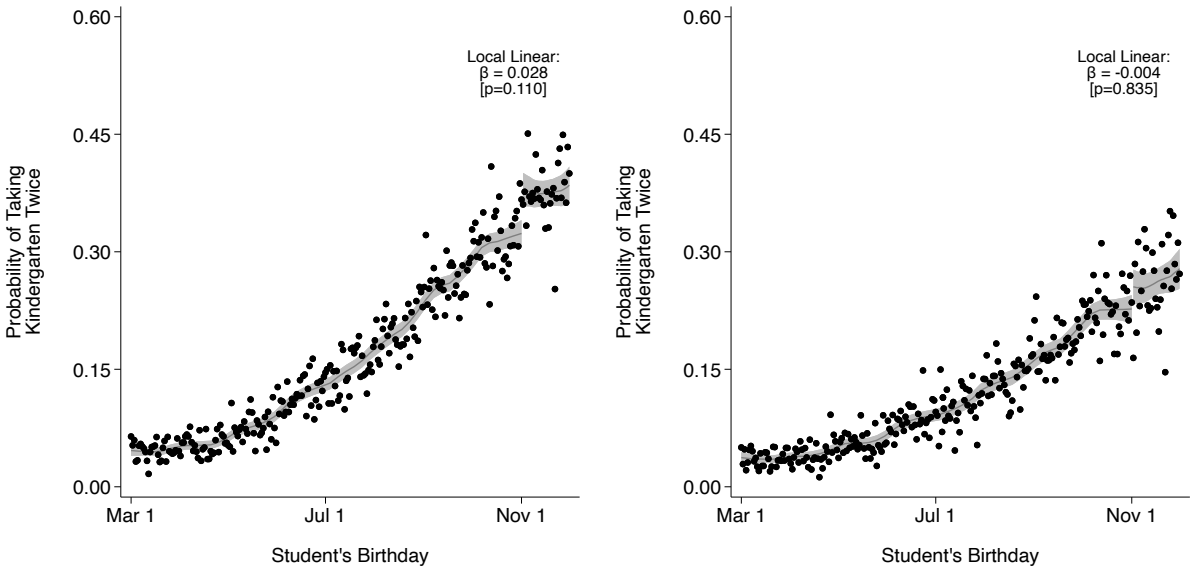
Table E.4: Treatment Effects Constant Dropping Possible Problems

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

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

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

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



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

and well estimated—both for the whole population and for the relevant subpopulations.

E.2 Testing Parallel Trends

Above I demonstrated that the effects estimated at the different cutoffs are significantly different from one another for students from higher-income families. Interpreting these differences as positive selection on gains hinges on the parallel trends assumption proposed in Section 2 that for a given type of child (characterized by u and x) changes in scores over birthdays would have the same slope for those who start as for those who wait. If this parallel trends assumption is not met, then differences estimated at the different cutoffs could be attributable to differences between the compliers at each cutoff or to heterogeneity in effects over birthdays. This subsection evaluates on testable implication of this parallel trends assumption.

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

grade at age nine). Unlike potential third-grade test scores, test scores in third grade and fourth grade are both directly observable.

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

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

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

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

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

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

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

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

slope between grades. This bounding logic is similar to the approach of Rambachan and Roth (2022) for bounding differences in differences estimators for the largest plausible deviation from parallel trends.

E.3 Bounding the Effect on Always Takers

E.3.1 Monotonic Selection

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

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

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

As we know that $\mu_{0,ec}(r_1) < \mu_{0,rc}(r_1)$, the bite of this assumption is in bounding $\mu_{0,at}(r_1)$ below $\mu_{0,ec}(r_1)$.

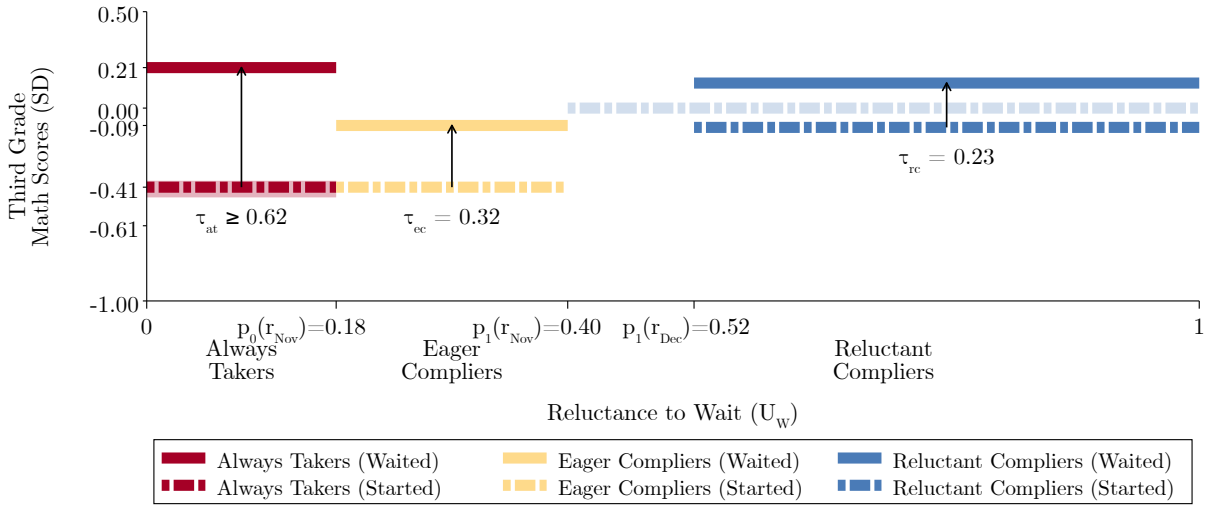
This inequality is not strong enough to identify a point estimate of the τ_{at}, but it can identify an informative bound on the average treatment effect for always takers:

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

Because I assume $\mu_{0,at} \leq \mu_{0,ec}$, the statistic $\tilde{\tau}_{at} = 0.64$ is a lower bound on τ_{at}[43] Because $\tilde{\tau}_{at} > 0$, it is an informative bound.[44]

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

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



Note: This figure graphically illustrates the smallest possible effect on always takers under monotonicity. This bound is recoverable from the average of student outcomes by intervention assignment (whether assigned wait to enter kindergarten) and treatment status (whether actually waited) and the auxiliary assumption of weakly monotonic expected untreated outcomes. The sample is comprised of 15,081 students who meet the following criteria: enter kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turn five within thirty days of November 1, 2013; and take standard third-grade math tests. Average outcomes for treated and untreated compliers are backed out of observed data and choice probabilities. For bootstrapped standard errors see table 5

To test the null of effect homogeneity between always takers and eager compliers under this weaker assumption I use a test for heterogeneity from in Kowalski (2022a). This test determines whether the eager-

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

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

complier LATE, τ_{ec}, falls within the range bounded by $\tilde{\tau}_{at}$. The one-tailed test statistic returns a 1 if it rejects treatment effect homogeneity (i.e., whether τ_{ec} is outside of the bounds from $\tilde{\tau}_{at}$) and a 0 otherwise:

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

Again for inference I use a nonparametric block bootstrap and report the percent of failures to reject as the p-value. Under this second assumption I reject homogeneity at the $p = 0.011$ level. Full results are reported in Table 5.

F. Appendix Tables and Figures

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

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

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

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

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

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

Table F.8: Early Elementary School Outcomes By Subgroup

Panel A: All Students	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		20.3%	77.3%			
Special Education (K)	0.099	0.170	0.069	0.086	0.101***	-0.018***
	(0.001)	(0.006)	(0.003)	(0.002)	(0.008)	(0.004)
Special Education (3)	0.147	0.210	0.114	0.138	0.096***	-0.025***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.009)	(0.005)
No Third Grade Math Score	0.110	0.154	0.083	0.116	0.071***	-0.033***
	(0.001)	(0.006)	(0.003)	(0.003)	(0.008)	(0.004)
Accommodated Test in Third Grade	0.008	0.013	0.006	0.006	0.008***	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Panel B: Low-SES Students	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		13.1%	85.0%			
Special Education (K)	0.116	0.279	0.078	0.099	0.201***	-0.022***
	(0.002)	(0.014)	(0.004)	(0.004)	(0.016)	(0.006)
Special Education (3)	0.178	0.330	0.137	0.171	0.193***	-0.035***
	(0.002)	(0.015)	(0.005)	(0.004)	(0.018)	(0.007)
No Third Grade Math Score	0.115	0.223	0.085	0.119	0.138***	-0.035***
	(0.001)	(0.012)	(0.004)	(0.004)	(0.015)	(0.006)
Accommodated Test in Third Grade	0.012	0.033	0.008	0.009	0.024***	-0.001
Panel C: Higher-SES Students	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		26.4%	71.0%			
Special Education (K)	0.088	0.128	0.059	0.074	0.069***	-0.015***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.010)	(0.005)
Special Education (3)	0.123	0.167	0.089	0.108	0.078***	-0.019***
	(0.001)	(0.007)	(0.005)	(0.004)	(0.011)	(0.006)
No Third Grade Math Score	0.105	0.127	0.080	0.110	0.047***	-0.030***
	(0.002)	(0.006)	(0.004)	(0.004)	(0.009)	(0.006)
Accommodated Test in Third Grade	0.005	0.006	0.003	0.003	0.003	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Panel D: Black	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		7.7%	87.6%			
Special Education (K)	0.091	0.238	0.065	0.082	0.174***	-0.017***
	(0.003)	(0.022)	(0.005)	(0.005)	(0.024)	(0.007)
Special Education (3)	0.153	0.317	0.115	0.161	0.202***	-0.046***
	(0.002)	(0.025)	(0.006)	(0.006)	(0.028)	(0.009)
No Third Grade Math Score	0.127	0.265	0.100	0.131	0.165***	-0.031***
	(0.002)	(0.022)	(0.005)	(0.005)	(0.024)	(0.008)
Accommodated Test in Third Grade	0.011	0.031	0.009	0.010	0.023**	-0.002
	(0.001)	(0.009)	(0.002)	(0.002)	(0.010)	(0.002)
Panel E: White	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		25.4%	72.8%			
Special Education (K)	0.105	0.161	0.072	0.093	0.088***	-0.020***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.009)	(0.005)
Special Education (3)	0.148	0.194	0.115	0.132	0.080***	-0.017***
	(0.002)	(0.007)	(0.005)	(0.004)	(0.010)	(0.006)
No Third Grade Math Score	0.094	0.128	0.066	0.098	0.061***	-0.031***
	(0.001)	(0.006)	(0.004)	(0.003)	(0.008)	(0.005)
Accommodated Test in Third Grade	0.008	0.012	0.006	0.004	0.006**	0.001
	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)
Panel F: Girls	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		16.6%	80.8%			
Special Education (K)	0.065	0.121	0.047	0.059	0.074***	-0.012***
	(0.001)	(0.008)	(0.003)	(0.003)	(0.010)	(0.004)
Special Education (3)	0.101	0.154	0.081	0.099	0.073***	-0.018***
	(0.001)	(0.009)	(0.004)	(0.004)	(0.011)	(0.006)
No Third Grade Math Score	0.106	0.163	0.077	0.114	0.086***	-0.037***
	(0.001)	(0.009)	(0.004)	(0.004)	(0.011)	(0.005)
Accommodated Test in Third Grade	0.005	0.010	0.004	0.003	0.006**	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)
Panel G: Boys	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		23.8%	74.0%			
Special Education (K)	0.132	0.203	0.092	0.115	0.111***	-0.023***
	(0.002)	(0.009)	(0.005)	(0.004)	(0.012)	(0.006)
Special Education (3)	0.191	0.246	0.149	0.180	0.097***	-0.031***
	(0.002)	(0.010)	(0.006)	(0.005)	(0.014)	(0.007)
No Third Grade Math Score	0.114	0.148	0.089	0.118	0.059***	-0.029***
	(0.001)	(0.007)	(0.004)	(0.004)	(0.009)	(0.006)
Accommodated Test in Third Grade	0.011	0.016	0.008	0.009	0.008**	-0.001
	(0.000)	(0.003)	(0.002)	(0.001)	(0.004)	(0.002)

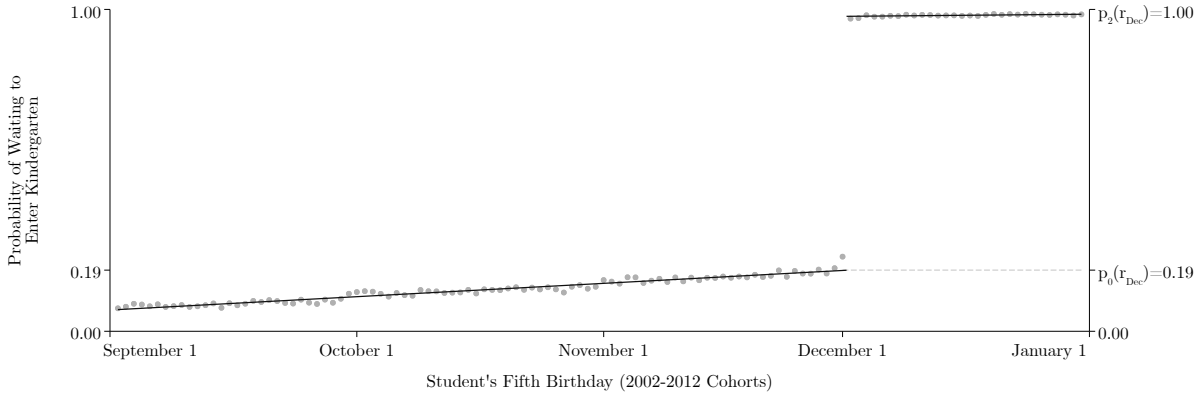
Note: This sample shows the average near-term outcomes for always takers, eager compliers, and reluctant compliers. The sample is comprised of 15,990 students who started kindergarten in Michigan public schools in 2002-13 and turn five within 90 days of December 1. Note that in this table I do not restrict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

Table F.9: Different Measures of ATEs

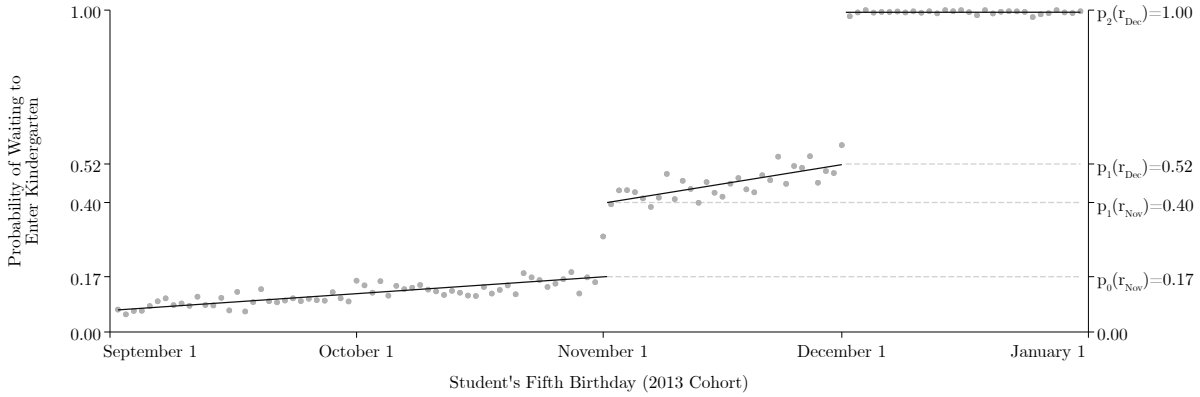
	Reluctant Complier Effect	Eager Complier Effect	Always Taker Effect	Average Effect (RC Bound)	Average Effect (EC Bound)
All Students	0.230***	0.333***	0.838***	0.361***	0.373***
	(0.071)	(0.134)	(0.198)	(0.082)	(0.090)
Higher-SES Students	0.156*	0.596***	1.179***	0.481***	0.550***
	(0.084)	(0.238)	(0.302)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

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

Figure F.6: Visualization of Policy Change



(a) 2002-2012 Cohorts: Only December Cutoff



(b) 2013 Cohort: November and December Cutoffs

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