How the Timing of Grade Retention Affects Outcomes

Author(s): Jane Cooley Fruehwirth, Salvador Navarro and Yuya Takahashi

Source: *Journal of Labor Economics*, October 2016, Vol. 34, No. 4 (October 2016), pp. 979-1021

Published by: The University of Chicago Press on behalf of the Society of Labor Economists and the NORC at the University of Chicago

Stable URL: https://www.jstor.org/stable/10.2307/26553232

How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects

Jane Cooley Fruehwirth, *University of North Carolina at Chapel Hill*

Salvador Navarro, *University of Western Ontario*

Yuya Takahashi, *Johns Hopkins University*

We show how the effect of grade retention varies by abilities, by timing of retention, and as time since retention elapses. Existing studies of grade retention are not well equipped to deal with the possibility that students retained at different grades differ in unobservable abilities (dynamic selection) and the effects of retention also vary by the student's abilities and the timing of retention. We extend existing factor analytic methods for identifying treatment effects in such settings. Using ECLS-K data, we find evidence of dynamic selection into retention and of heterogeneous effects of retention by grade and unobservable abilities.

I. Introduction

Grade retention (or grade repetition) is a common and controversial practice in many countries. In Germany, 9% of 15-year-olds report being

We thank Pat Bajari, Anirban Basu, Gary Becker, William "Buz" Brock, Juan Esteban Carranza, Flavio Cunha, Steven Durlauf, Jim Heckman, Lynne Heckman, Han Hong, Caroline Hoxby, Ju Hyun Kim, Kevin Lang, David Lee, David Meltzer, Dan Millimet, Jim Walker, Ken Wolpin, participants at numerous seminars, and

[*Journal of Labor Economics*, 2016, vol. 34, no. 4]
 0734-306X/2016/3404-0005$10.00
Submitted May 14, 2013; Accepted March 23, 2015; Electronically published August 3, 2016

retained in primary school, and as many as 18% in France do so (Eurydice 2011). In the United States, about 10% of students are retained between kindergarten and eighth grade (approximately age 14; NCES 2009). With the federal reforms initiated by the No Child Left Behind Act (NCLB) of 2001 and associated national rhetoric emphasizing an end to social promotion, the practice of grade retention is facing new scrutiny as a potential policy to help bring students up to proficiency levels in the United States. This is somewhat surprising given that the empirical literature often finds negative effects of retention and at best provides mixed evidence of its effectiveness in improving student outcomes (e.g., Holmes 1989; Jimerson 2001; Allen et al. 2009). In this paper, we provide evidence on how the effect of grade retention varies by age, the time the student is retained, and unobserved abilities. Understanding these types of heterogeneity in responses to grade retention is central to informing policy, both concerning who would benefit from retention and optimal timing. For instance, student accountability policies based on retention may vary in effectiveness depending on the average ability of the students who are retained (as determined by the cutoff for passing the exam) and whether the policy applies to students in early or later grades. Furthermore, to the extent that grade retention has less severe effects on socioemotional development in young children, early retention may lead to better outcomes than late retention.

We develop a simple framework to estimate time-varying treatment effects. As in the static treatment effect setting, the key challenge is dealing with selection on unobservables. However, in our context, the selection problem is further complicated because different types of students might select (or be selected) into treatment over time, that is, there may be dynamic selection. This is likely because the pool of potential retainees changes as students age and/or the rules for retention might also vary over time. Our method for controlling for dynamic selection can be understood as a hybrid between a control function and a generalization of the fixed effect approach. We assume that a "low" dimensional set of unobservables affects both selection into treatment and the outcome of treatment. This strategy effectively places restrictions on the covariances between unobservables in the outcome and selection equations, a generalization of the semiparametric factor structure of Carneiro, Hansen, and Heckman (2003); see also Bonhomme and Robin (2010) and Cunha, Heckman, and Schennach (2010) for recent

especially John Kennan, Karl Scholz, Jeff Smith, and Chris Taber for providing helpful comments. Navarro's work was supported by the Institute for Research on Poverty at the University of Wisconsin–Madison and SSHRC in Canada. Fruehwirth's work was supported by the British Academy Mid-Career Fellowship and the Leverhulme Trust's Leverhulme Prize. Contact the corresponding author, Salvador Navarro, at s.navarro.lozano@gmail.com. Information concerning access to the data used in this article is available as supplementary material online.

developments). It is a control function approach because we use information from the selection equation to help control for selection, so that the same unobserved abilities affect both test scores and the probability of being retained. Identification is further aided by the use of exclusion restrictions (retention policies), variables that affect a student's selection into retention but do not affect their outcomes directly.

This is akin to a fixed effects approach, in that we assume unobservable ability is fixed, but this generalizes because we can control for multiple dimensions of ability. This is particularly important when different types of unobservable ability, such as behavioral and cognitive abilities, play important and distinctive roles in determining both selection into treatment and outcomes, as in our application. In addition, unlike the fixed effect model, which assumes that fixed ability affects outcomes and selection in the same way over time, our framework allows the effects of unobservable abilities to be time-varying. This proves to be important in estimating how the effect of retention varies across grades.

Our paper contributes to a broad literature on the effects of grade retention. Most of this literature relies on the assumption of selection on observables (see Jimerson [2001] for an overview). More recent literature recognizes the importance of unobservables in confounding the effect of grade retention and has developed innovative approaches to deal with selection on unobservables. Gary-Bobo, Gousse, and Robin (2013) is most closely related to our paper, in that its authors apply a factor-analytic model to the question of grade retention, but it does not allow for time-varying treatment effects.[1] Other studies use both regression discontinuity and instrumental variables to deal with selection on unobservables (Eide and Showalter 2001; Fertig 2004; Jacob and Lefgren 2004, 2009; Greene and Winters 2007; Brodaty, Gary-Bobo, and Prieto 2008; Manacorda 2012). For instance, Jacob and Lefgren (2009) and Manacorda (2012) use a regression discontinuity approach that exploits a clear policy (a test performance threshold in the case of Jacob and Lefgren [2004, 2009] and a rule based on days absent in the case of Manacorda [2012]). Considering the case of the achievement rule, the treatment effect of grade retention is estimated by comparing students just above to students just below the threshold. These studies are also interested in heterogeneity by time at which the student is retained and compare estimated treatment effects across different grades where performance thresholds apply. However, when treatment effects vary by the unobservable ability of students, these comparisons are confounded by dynamic selection: the marginal and average unobserved ability of students retained in these

[1] Another notable difference is that our extended factor model accounts for multiple dimensions of ability (cognitive, behavioral, etc.), while Gary-Bobo et al. (2013) use single-dimensional finite ordered types.

different grades is likely to vary because of prior retention decisions and also because of different performance thresholds. The key contribution of our method is to separate out this time-varying treatment effect from dynamic selection.

A second contribution of our method relative to the existing literature that uses instrumental variables (IV) or regression discontinuity is that we are able to show how the treatment effects vary by the unobserved abilities of students and to estimate different average treatment effects, such as treatment on the treated, rather than local average treatment effects. Because of heterogeneity in retention policies across states, we have sufficient support to identify the effect of grade retention even for the lowest-ability students, who are often the target of policy, rather than just the marginal students. We can also provide insight into how treatment effects might differ across students depending on whether they are of low cognitive or behavioral ability. Because we have considerable overlap in the distributions of abilities across retention statuses (even students with above-median cognitive test scores are retained in our data), we can also consider effects on relatively high-ability students.

Methodologically, our paper also contributes to a relatively sparse literature on the estimation of time-varying treatment effects. Ham and LaLonde (1996) and Abbring and Van den Berg (2003) provide other useful approaches to analyzing treatment effects in dynamic models. Whereas they rely on the proportional hazards assumption, our model supports more general forms of treatment heterogeneity than in either Abbring and Van den Berg (2003), where treatment heterogeneity can be allowed at the expense of ruling out the endogenously selected time-at-treatment to affect outcomes, or Ham and LaLonde (1996), where treatment effects are homogeneous. Cellini, Ferreira, and Rothstein (2010) and Fruehwirth and Traczynski (2013) use a dynamic regression discontinuity approach, but these approaches deal with a more limited form of selection and generally are not informative about treatment effects away from the threshold, a central contribution of our approach. Rokkanen (2014) and Dong and Lewbel (2015) provide possible extensions to the regression discontinuity method for estimating treatment effects away from the threshold, but this would still only apply to specialized settings where the selection rule is known. An advantage of our method is that it generalizes to a majority of settings, where the selection rule for retention is unknown. For instance, students may be retained through parental request, because they have low behavioral ability, or because they underperform on a cognitive exam.

Our approach to modeling time-varying treatments is close to that in Heckman and Navarro (2007). However, our focus is substantively different, namely, on how factor analytic methods can aid in identification and interpretation of time-varying treatment effects. This builds on a burgeoning literature that shows how factor models can be useful for identifying treatment effects, primarily in a static context (e.g., Urzua 2013; Rokkanen

2014). We provide a further methodological contribution in generalizing the factor structure results used in other settings (Carneiro et al. 2003; Bonhomme and Robin 2010). By employing the information contained in higher-order moments, our factor model is less "data hungry" than standard models and can be applied to settings where few measurements (e.g., test scores) are available, but with the additional assumption that the factors are distributed asymmetrically (see Bonhomme and Robin [2009] for a similar approach). Furthermore, an important aspect of our model, where we are considering time-varying treatment effects, is to relax the standard assumption of uncorrelated unobservable shocks that enter each equation in each period by exploiting the panel aspect of our data to allow new unobservable factors every period (unobserved persistent shocks). This is important because the event of retention in a given year could be correlated with other negative shocks (such as parental divorce) that simultaneously affect test outcomes.

We evaluate the effect of retention on achievement using data from the Early Childhood Longitudinal Study of Kindergartners (ECLS-K). We find that students who are retained in kindergarten would have performed as much as 27% higher in the next year if they had not been retained. We also find that the initial losses to achievement diminish over time. By the end of our data, when students are approximately age 11, eliminating grade retention raises achievement by as much as 7% for students who were retained in prior years. This means that these retained students learn 7% less by age 11 than they would have learned if they had not been retained. As we discuss further below, a somewhat surprising finding is that the treatment effect of kindergarten retention is positive for the average untreated student in the long run, whereas it is negative for the average treated student. We provide evidence that the positive average treatment effect arises because higher-ability students, if retained, would receive more parental and school resources as a consequence of retention than lower-ability retained students.

The paper proceeds as follows. In Section II, we describe the basic framework and define dynamic treatment effects for the dynamic case. In Section IV, we specialize the framework to our proposed factor structure. We show that the model is semiparametrically identified. We describe our estimation strategy in Section V. Data and results are discussed in Sections III and VI.

II. The Framework

Below we outline a simple framework for evaluating the effect of grade retention that permits heterogeneous treatment effects based on ability and timing of retention and applies to a general context where the selection rule is unknown.

Let $t = 1, 2, ..., \bar{t}$ index calendar time and $i = 1, ..., I$ index the individual. Since we allow for students to be retained at different times, we define a random variable that indicates the grade in which a student is retained, $R_i \in \{1, 2, ..., \overline{R} - 1, \overline{R}, \infty\}$, where $\overline{R} \leq \bar{t}$ allows for the possibility that students

may be retained only up to a certain time period or grade. We assume that the student is retained at most once. Our data follow a single cohort of kindergarteners across time, so that $R_i = 1$ denotes that a student is retained in kindergarten, and so forth. We adopt the convention of letting $R_i = \infty$ for the "never" treated state where we do not observe a student being retained in the sample period.

A. Outcomes and Selection Rule

The outcome of interest, math and reading test scores at time t for a student i who receives treatment at time r, is denoted by $Y_i(t, r)$. For notational simplicity, we keep all conditioning on covariates, observable school and student characteristics, implicit. Finally, we define a random variable $D_i(r)$ that takes value 1 if an individual is retained at time r and 0 otherwise. For individual i, the observed outcome in period t will be given by

$$Y_i(t) = \sum_{r=1}^{\bar{R}} D_i(r)[Y_i(t, r) - Y_i(t, \infty)] + Y_i(t, \infty) . \quad (1)$$

While the standard case only has the treated and untreated potential states, we have the untreated, the treated at time 1, the treated at time 2, and so forth. There is no single effect of retention, but rather an effect of retention in kindergarten, in first grade, and so forth. Furthermore, there is no single effect of retention in kindergarten, for example, as the effects depend on the time elapsed since retention.

Following Abbring and Van den Berg (2003), we also impose that:

ASSUMPTION 1. $Y_i(t, r) = Y_i(t, \infty) = Y_i(t)$ for $r \geq t$.

This assumption rules out that potential outcomes differ because in the future treatment times will be different. This means, for example, that after conditioning on all prior information, the fact that a student will be retained in second grade does not directly affect his/her performance in first grade. While Abbring and Van den Berg refer to this as the *no anticipations* assumption, this should not be confused with the assumption that individuals are not forward-looking. Assumption 1 does not rule out that individuals may predict that they are more likely to get treated at a particular time r (i.e., have some anticipation as to treatment time).[2]

[2] The assumption does rule out that after conditioning on the information available at the pre-r period of interest t, the actual event of getting treated at time r has an effect on pre–time r outcomes. It is in this sense that it is closer to a "no perfect foresight" assumption, although this is not necessary for assumption 1 to hold. We can accommodate cases in which assumption 1 does not hold, but we keep the assumption for simplicity. See Abbring and Van den Berg (2003) and Heckman and Navarro (2007) for a discussion.

We decompose outcomes into a mean component and an individual-specific component,

$$Y_i(t,r) = \Phi(t,r) + \epsilon_i(t,r), \tag{2}$$

where, because of assumption 1, we impose $\Phi(t,r) = \Phi(t)$ and $\epsilon_i(t,r) = \epsilon_i(t)$ if $r \geq t$. This decomposition permits us to write the observed outcome (test score) in period t as

$$Y_i(t) = \Phi(t,\infty) + \epsilon_i(t,\infty) + \sum_{r=1}^{\min\{t,\bar{R}\}} D_i(r)(\Phi(t,r) - \Phi(t,\infty)) + \sum_{r=1}^{\min\{t,\bar{R}\}} D_i(r)(\epsilon_i(t,r) - \epsilon_i(t,\infty)).$$

This is similar to the typical potential outcomes framework, except that now we have multiple periods where students could be retained, and so we have to allow for all potential treatment times. The first component describes the outcome for those not retained. The second and third components describe, respectively, the mean and individual-specific treatment effect of retention at a given grade (relative to not being retained). The relevance of this representation becomes clearer in the discussion of identification and estimation.

In most cases, the decision to retain a student is not clearly defined, but rather is the result of a complex process involving many actors, including teachers, principals, and parents. We thus model selection in a reduced-form way, such that the treatment-time-specific index is $V_i(r) = \lambda(r) + U_i(r)$ for $r \in \{1, 2, \ldots, \overline{R} - 1, \overline{R}\}$, and treatment time is selected according to

$$\begin{aligned} D_i(R_i) &= 1\left(V_i(R_i) > 0 \mid \{V_i(r) < 0\}_{r=1}^{R_i-1}\right) \\ &= 1\left(V_i(R_i) > 0 \mid \{D_i(r) = 0\}_{r=1}^{R_i-1}\right), \end{aligned}$$

where $\mathbf{1}(a)$ is an indicator function that takes value 1 if a is true and 0 otherwise, and where $R_i = \infty$ if $\{V_i(r) < 0\}_{r=1}^{\overline{R}}$. The selection process is dynamic in the sense that today's choice to retain a student depends on yesterday's choice: treatment time r can only be selected if treatment has not been taken before.

This framework can be thought of as a midpoint between the standard static treatment literature that does not model the selection process explicitly and a fully specified structural dynamic discrete choice model.[3] At

[3] Our selection model is consistent with the usual threshold-crossing or reservation-value decision rules that frequently arise from complex dynamic decision problems. Cunha, Heckman, and Navarro (2007) provide conditions under which structural dynamic discrete choice models can be represented by a reduced form approximation as above.

the same time, the selection process we propose is consistent with, for example, the commonly employed test score thresholds for whether a student should repeat a grade. This threshold could be individual specific if schools use relative comparisons or take into account extenuating circumstances for individual students. Our selection process applies whether we observe the scores used for the decision (as in Jacob and Lefgren 2004) or not. For example, if the jth test score $Y_{i,j}(t)$ (whether observed by the econometrician or not) is used to decide who to retain, and the threshold μ_i is individual specific, we would have

$$\begin{aligned} V_i(t) &= \lambda(t) + U_i(t) \\ &= -Y_{i,j}(t) + \mu_i \\ &= -\Phi_j(t) - \epsilon_{i,j}(t) + \mu_i, \end{aligned} \tag{3}$$

where $\lambda(t) = -\Phi_j(t)$ and $U_i(t) = -\epsilon_{i,j}(t) + \mu_i$. Clearly, thresholds based on combinations of different test scores would also be consistent with our specification.

B. Defining Treatment Effects

Because both the timing of treatment and the time elapsed since treatment may matter, there are many possible individual treatment effects. A particular parameter of interest is

$$\begin{aligned} \Delta_i^1(t, r, r') &= Y_i(t, r) - Y_i(t, r') \\ &= \Phi(t, r) - \Phi(t, r') + \epsilon_i(t, r) - \epsilon_i(t, r'), \end{aligned}$$

which measures the effect at period t of receiving treatment at time r versus receiving treatment at time r'. An example would be the difference in test scores at age 11 for a student if he/she repeats first grade versus if he/she repeats third grade. If we let $r' = \infty$, this parameter would measure the effect at t of receiving treatment at time r versus not receiving treatment at all.

Because of the multiplicity of treatments available, we can define many more mean treatment parameters than in the static binary case, like the average effect of receiving treatment at r versus receiving treatment at r',

$$\text{ATE}(t, r, r') = E(Y(t, r) - Y(t, r')) = \Phi(t, r) - \Phi(t, r')$$

or the effect of treatment at r versus treatment at r' for people who are actually treated at time $R_i = r''$,

$$\text{TT}(t, r, r', r'') = E(Y(t, r) - Y(t, r') \mid R_i = r''),$$

and so forth. For instance, we may want to know the return to retaining students in kindergarten who were actually retained in first grade. One inter-

esting issue in comparing treatment effects is that we cannot hold both grade and age fixed. In this paper, we hold age fixed because that is what is permitted with the data, as discussed further below. In the case of being treated versus not, the treatment effect of not being retained involves the student being exposed to an extra grade's worth of material. In comparing the relative effects of two possible treatment times, such as being retained in kindergarten versus first grade, the student is exposed to the same amount of material since the treatment comparison is done at the same grade (e.g., third grade).

III. Data

We use the ECLS-K, a nationally representative survey of kindergartners in school year 1998–99, to study the effect of grade retention. It follows the students as they progress through school, with follow-up surveys in the 1999–2000, 2001–2, and 2003–4 school years. A benefit of these data is that we observe the history of a student's schooling beginning at kindergarten, and it covers the earlier years when retention is relatively more common. Roughly 10% of our sample is retained between kindergarten and fourth grade. We restrict the sample to students who were retained at most once, did not skip grades, and were taking kindergarten for the first time in 1998–99.[4] Because of the nature of the survey, we are able to form three different retention indicators: kindergarten, early (first or second grades), and late (third or fourth grades).[5] That is, our dynamic treatment time indicator takes values $R_i = 1, 2, 3, \infty$, where $R_i = \infty$ means the student is never retained, $R_i = 1$ that he/she is retained in kindergarten, $R_i = 2$ that he/she is retained early, and $R_i = 3$ that he/she is retained late.

Each year of the ECLS-K includes cognitive tests measuring students' science, reading, and math skills.[6] We focus primarily on the effect of retention at different grades on the math and reading tests, using the log of the item response theory (IRT) scores. The ECLS-K also includes teacher ratings on students' behavioral and social skills—the approaches to learning, self-control, and interpersonal skills components of the Social Rating Scale

[4] The number of students we observe being retained twice in the raw data is about .3% of the sample. After restricting to the sample with the necessary set of covariates, this number would be even smaller. We lose about 100 students in the restricted sample when we drop students who are taking kindergarten for the second time in the base year, or about 1% of our restricted sample. Including them does not significantly change our estimated effects of retention.

[5] We can separate early and late into the four grades at which retention takes place, but only for less than half of the sample, and we already lose a significant amount of data because of attrition, as shown below.

[6] In the first two periods, students are given a general knowledge test, rather than a science test, which measures science skills. However, the science and the general knowledge tests are not directly comparable.

(SRS). We use these together with the cognitive tests in order to identify the different components of ability, as described below in Section IV.

A logical difficulty in evaluating the effect of grade retention is that it is impossible to hold both the grade and the age fixed when determining the treatment effect of retention. Depending on the policy question of interest, it may be more appropriate to focus on measuring effects holding grade fixed or holding age fixed. The effect holding grade fixed addresses, for instance, whether a student learns more by the end of fifth grade than he/she would have if he/she had not repeated fourth grade. This attributes maturation (or age) effects to the estimated effect of retention. Alternatively, holding age fixed measures whether a student learns more, say, by age 11 if he/she repeats fourth grade than he/she would have if he/she had been promoted to the fifth grade and exposed to new material. We focus on the effect of retention holding age fixed.[7] This is in part because the tests used by ECLS-K are designed to measure cognitive development as opposed to grade-specific knowledge, but also because we do not have the data to estimate the effect holding grade fixed. For instance, if a student is retained in kindergarten, we observe his/her achievement twice in kindergarten and then in second and fourth grade but not in third grade and fifth grade, which is what we observe for nonretained students. Arguably, estimates holding age fixed provide a more conservative estimate of the potential benefits of grade retention, given that we are comparing students who have not been exposed to an extra year of material.

The ECLS-K contains a very rich set of covariates. We use characteristics of the student, the family, the class, and the school as controls in our model. Class and teacher characteristics are taken from teacher surveys.[8] School administrator surveys provide information about the school characteristics, and parent surveys provide information about the family.

Table 1 shows descriptive statistics for the covariates we include in all our equations for the first year of the survey (1998–99) in columns 1–3. In order to include as many observations as possible, we include in the sample students who have any test score measure in the first year and the full set of conditioning covariates. Thus, the number of observations differs across test scores and covariates. A potentially important concern with a panel study of this type is nonrandom sample attrition. As shown in table 1, the number of observations decreases substantially from 7,832 in the base year to 2,106 in the last year. Column 5 of table 1 shows the mean 1998–99 characteristics

[7] This is also the focus of Jacob and Lefgren (2004) and a number of the higher-quality studies in the literature, as surveyed by Allen et al. (2009).

[8] For the 2003–4 school year, both math/science and reading teachers fill out surveys, resulting in potentially different classroom and teacher characteristics for math/science and reading. We use the relevant classroom measures for each test in estimating the outcome equations.

Table 1
Summary Statistics

	Value of Variables in 1998–99 School Year for Observations Included in:					
	1998–99 School Year			2003–4 School Year		
Variable	Observations	Mean	SD	Observations	Mean	SD
General test score	7,549	3.09	.35	2,078	3.14	.33
Reading test score	7,608	3.36	.28	2,078	3.39	.27
Math test score	7,794	3.10	.36	2,101	3.14	.35
Approach to learning	7,829	.05	.98	2,104	.13	.95
Self-control	7,808	.03	.97	2,097	.11	.94
Interpersonal skills	7,782	.02	.98	2,095	.09	.96
Male	7,832	.50	.50	2,106	.49	.50
White	7,832	.65	.48	2,106	.77	.42
Black	7,832	.12	.32	2,106	.07	.26
Hispanic	7,832	.14	.34	2,106	.09	.28
Body mass index	7,832	16.25	2.13	2,106	16.21	2.10
Age	7,832	5.62	.34	2,106	5.63	.34
Number of siblings	7,832	1.42	1.11	2,106	1.41	1.07
Socioeconomic Status Index	7,832	.10	.78	2,106	.20	.74
Attended full-time kindergarten	7,832	.58	.49	2,106	.52	.50
TV rule at home	7,832	.89	.32	2,106	.89	.31
Mother not in household	7,832	.01	.11	2,106	.01	.11
Father not in household	7,832	.17	.37	2,106	.12	.32
Number of books at home	7,832	80.54	60.75	2,106	88.76	60.23
Minority students in school between (1%, 5%)	7,832	.20	.40	2,106	.20	.40
Minority students in school between (5%, 10%)	7,832	.15	.36	2,106	.12	.33
Minority students in school between (10%, 25%)	7,832	.10	.30	2,106	.05	.22
Minority students in school > 25%	7,832	.16	.36	2,106	.09	.29
Public school	7,832	.78	.42	2,106	.73	.44
TT1 funds received by school	7,832	.62	.49	2,106	.63	.48
Crime a problem	7,832	.46	.58	2,106	.36	.52
Students bring weapons	7,832	.16	.37	2,106	.13	.34
Children or teachers physically attacked	7,832	.36	.48	2,106	.35	.48
Security measures in school	7,832	.55	.50	2,106	.58	.49
Parents involved in school activities	7,832	2.97	.90	2,106	3.10	.83
Teacher has a master's degree	7,832	.35	.48	2,106	.34	.48
Teacher experience	7,832	14.31	9.03	2,106	14.39	8.97
Student's class size	7,832	20.40	5.00	2,106	19.89	4.80
Teacher's rating of class behavior	7,832	1.56	.78	2,106	1.52	.77
Minority students in class between (1%, 5%)	7,832	.08	.26	2,106	.09	.29
Minority students in class between (5%, 10%)	7,832	.13	.33	2,106	.16	.36
Minority students in class between (10%, 25%)	7,832	.18	.39	2,106	.18	.38
Minority students in class > 25%	7,832	.42	.49	2,106	.28	.45

SOURCE.—ECLS-K Longitudinal Kindergarten–Fifth Grade Public-Use Data File.

NOTE.—For our counterfactual analyses, we only use data on students whose covariates and retention history are observable (i.e., not missing) for all time periods. Thus, we end up with fewer observations at the 2003–4 school year.

for students who are still in the sample in 2003–4 (the last year of the survey that we use for estimation). Comparing summary statistics, we see suggestive evidence of nonrandom attrition. In Section IV.C, we discuss how we control for nonrandom attrition. Table 2 breaks out summary statistics by retention statuses and shows that students who are retained have lower test scores and are more disadvantaged than those who are not retained. Furthermore, observable characteristics differ across retention statuses, suggesting that unobservable characteristics may differ as well, that is, that dynamic selection might be a concern.

Before discussing our identification strategy, we first perform some baseline OLS regressions that indicate that dynamic selection and/or time-varying treatment effects are likely to be important in our data. To test for dynamic selection, we regress the kindergarten cognitive tests, which took place prior to any retention decisions, on period-specific indicators of whether the student is retained in the future. We also control for covariates related to the student, his family, school, and class, as described in table 1 above. Column 1 of table 3 presents results for reading and math in panels A and B, respectively. Not surprisingly, students who will be retained have lower kindergarten test scores than those who will not be retained. Reading scores are 18% lower for kindergarten retainees and 20% and 12% lower for early and late retainees. Math scores are even more striking, 27%, 32%, and 22% lower for kindergarten, early retainees, and late retainees, respectively. Furthermore, *p*-values, reported at the bottom of the table, reject the joint test that the coefficients on being retained at different grades in the future are the same. These results suggest not only the presence of selection but also dynamic selection on cognitive test scores, that is, different types of students are being retained at different grades.

We show evidence that time-varying treatment effects are likely to be present by regressing test scores in the last sample period (2003–4 school year) on retention in different grades. As shown in column 2 of table 3, being retained is associated with worse outcomes than not being retained. The coefficients on the different retention statuses are also significantly different from each other. This is not direct evidence of time-varying treatment effects, since differences in the estimated effects across grades could be a result of time-varying treatment effects or a result of dynamic selection.

One way to begin to control for a static component of selection is to include various performance measures in kindergarten, that is, prior to any retention decisions taking place. Columns 3 and 4 control for kindergarten cognitive test scores and then behavioral scores. Consistent with the existence of selection, the negative effects of retention become smaller but do not disappear. For instance, the coefficient on kindergarten retention is cut in half for both reading and math, from −18% without initial test controls to −9% with test controls. Furthermore, we reject the formal test of equality of the effects for different retention times, again providing evidence

Table 2
Summary Statistics for Selected Variables by Retention Status (1998–99 School Year)

	Not Retained		Retained in Kindergarten		Retained Early		Retained Late	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
General test score	3.12	.33	2.85	.37	2.72	.33	2.78	.32
Reading test score	3.39	.27	3.13	.21	3.08	.18	3.15	.17
Math test score	3.14	.35	2.77	.32	2.67	.26	2.74	.25
Approach to learning	.12	.94	−.72	.99	−.91	.95	−.40	.98
Self-control	.06	.96	−.31	1.02	−.41	1.03	−.09	.93
Interpersonal skills	.06	.96	−.36	.95	−.53	1.00	−.21	1.01
Male	.49	.50	.66	.48	.63	.48	.54	.50
Black	.11	.31	.14	.35	.29	.46	.28	.45
Hispanic	.13	.34	.12	.32	.19	.39	.18	.39
Age	5.64	.34	5.39	.28	5.50	.32	5.52	.33
Attended full-time kindergarten	.57	.49	.62	.49	.61	.49	.72	.45
Number of siblings	1.39	1.08	1.65	1.27	1.80	1.41	1.52	1.25
Socioeconomic Status Index	.13	.77	−.12	.80	−.33	.69	−.54	.60
TV rule at home	.89	.31	.90	.30	.83	.37	.90	.31
Father not in household	.16	.37	.19	.39	.28	.45	.38	.49
Number of books at home	82.52	60.84	71.20	60.34	50.19	49.66	45.00	42.67
Minority students in school >25%	.15	.36	.16	.37	.27	.44	.38	.49
Public school	.77	.42	.73	.44	.91	.28	.93	.25
TT1 funds received by school	.62	.49	.61	.49	.76	.43	.79	.41
Teacher has a master's degree	.35	.48	.32	.47	.40	.49	.33	.47
Teacher experience	14.37	9.02	14.19	9.29	13.74	8.90	12.51	9.14
Student's class size	20.46	4.96	19.48	5.49	20.76	4.70	20.63	4.47
Minority students in class > 25%	.40	.49	.42	.50	.63	.48	.66	.48
Policy:								
Can be retained for immaturity	.76	.43	.78	.41	.72	.45	.68	.47
Can be retained at parents' request	.75	.43	.76	.43	.79	.41	.76	.43
Can be retained due to academic deficiencies	.88	.33	.83	.38	.91	.29	.88	.32
Can be retained any grade more than once	.10	.30	.13	.33	.14	.35	.15	.36
Can be retained more than once in elementary school	.35	.48	.30	.46	.43	.50	.50	.50
Can be tetained without parents' permission	.44	.50	.45	.50	.61	.49	.58	.50
Number of observations	7,038		255		288		87	

SOURCE.—ECLS-K Longitudinal Kindergarten–Fifth Grade Public-Use Data File

NOTE.—For our counterfactual analyses, we only use data on students whose covariates and retention history are observable (i.e., not missing) for all time periods. Thus, we end up with fewer observations at the 2003–4 school year. The last line lists the total number of usable observations (i.e., observations that contain at least one test/rating). Hence, the number of usable observations for any particular test/rating does not necessarily correspond to the number of observations in the last line. Notice that the last line does not sum to the total number of observations in table 1 (7,832). This is because we do not know every student's retention status. Regardless, these observations can still be used in period 1, when no selection has taken place.

Table 3
Evidence for Dynamic Selection and Treatment Effect

	Dependent Variable			
	Kindergarten Reading Score for 1998-99 School Year	Reading Score for 2003–4 School Year		
	(1)	(2)	(3)	(4)
A. Reading score:				
Retained in kindergarten	−.1775*	−.1791*	−.0948*	−.0926*
Retained early (1st or 2nd grade)	−.2014*	−.2306*	−.1450*	−.1374*
Retained late (3rd or 4th grade)	−.1222*	−.1192*	−.0498	−.0358
Student's characteristics	Yes	Yes	Yes	Yes
Family characteristics	Yes	Yes	Yes	Yes
School characteristics	Yes	Yes	Yes	Yes
Age and $(Age)^2$	Yes	Yes	Yes	Yes
Kindergarten cognitive tests		No	Yes	Yes
Kindergarten behavioral ratings		No	No	Yes
Number of observations	5,319	2,040	2,014	1,998
p-value for:				
KI = EA = LA	.003	.019	.026	.012
KI = EA	.189	.099	.079	.113
EA = LA	.001	.006	.009	.003
KI = LA	.028	.148	.192	.092
R^2	.312	.385	.530	.530
B. Math score:				
Retained in kindergarten	−.2735*	−.1804*	−.0727*	−.0889*
Retained early (1st or 2nd grade)	−.3172*	−.2450*	−.1463*	−.1396*
Retained late (3rd or 4th grade)	−.2240*	−.1697*	−.0875*	−.0387
Student's characteristics	Yes	Yes	Yes	Yes
Family characteristics	Yes	Yes	Yes	Yes
School characteristics	Yes	Yes	Yes	Yes
Age and $(Age)^2$	Yes	Yes	Yes	Yes
Kindergarten cognitive tests		No	Yes	Yes
Kindergarten behavioral ratings		No	No	Yes
Number of observations	5,462	2,043	2,017	1,998
p-value for:				
KI = EA = LA	.006	.094	.086	.012
KI = EA	.097	.071	.038	.076
EA = LA	.002	.079	.097	.004
KI = LA	.136	.813	.684	.141
R^2	.408	.357	.531	.522

NOTE.—Yes/No indicates if each group of variables is included as controls. KI, EA, and LA stand for the coefficient of the dummy variable for "retained in kindergarten," "retained early," and "retained late," respectively. The *p*-values are for the hypothesis of equality of coefficients on KI, EA, and LA.
* Statistically significant at the 5% level.

for potentially time-varying treatment effects. The same pattern holds for the other cognitive tests and behavioral measures. After including all initial test controls, retention in kindergarten is estimated to lower achievement by 9%, early retention by 14%, and late retention by only 4%, in both reading and math.

While this provides suggestive evidence of both time-varying treatment effects and dynamic selection, it is far from conclusive. The assumption that kindergarten test scores control for dynamic selection is a very restrictive one in that it assumes a static ability that determines whether one is retained in kindergarten, early, or late. In addition, test scores are noisy measures of true latent abilities; hence, using the kindergarten measures as controls may actually worsen the bias in the estimated treatment effects (Heckman and Navarro 2004). Furthermore, this illustrative analysis does not capture heterogeneous effects of treatment by student type, a central motivation of our paper.

The ECLS-K also has information on the schools' retention policies for the 1998–99, 1999–2000, and 2001–2 survey years, including whether the school has a policy that allows students to be retained in any grade (this policy only applies to grades after kindergarten), to be retained because of immaturity, to be retained at the parents' request, to be retained without parental authorization, to be retained multiple times, or to be retained multiple times in a given grade. As shown in table 2, retention policies vary considerably across schools and also to a lesser extent across retention statuses. In general, students who are retained early or late attend schools with more "liberal" retention policies than students who are not retained or who are retained in kindergarten. For instance, in the 1998–99 school year, 44% of schools in the nonretained sample permit retention without parental permission, compared to 61% and 58% for students who are retained early or late. Our identification strategy in Section V incorporates these variables by using them as exclusions, in that they do not directly determine the student's test score but they do affect the probability that a student repeats a grade.

We perform various tests to examine whether these are valid exclusions, focusing on the static setting, estimating the effect of kindergarten retention on 1999–2000 test scores, the year after retention. First, we perform a simple two-stage least squares regression (instrumenting for kindergarten retention with the retention policies) and find that they satisfy the test of overidentifying restrictions. We also try an alternative where we control for selection using a semiparametric control function approach (see Navarro [2008] for a description). If the exclusion restrictions are valid, they should only affect test scores through the selection process. Hence, we can test whether the retention policy variables are significant once we control for selection. To do this, we first estimate a probit of whether a student is retained in kindergarten on all the relevant covariates and the retention policy variables (the equivalent to the first stage in a 2SLS approach), and we find that the retention policy variables are jointly significant (p-value of .008). We then regress test scores on all the covariates, an indicator for whether the student was retained in kindergarten and a polynomial on the probability of selection (i.e., the control function). The terms involving the probability of selection are jointly significant in all cases (p-values of .033, .019, and .005

for reading, math, and general knowledge test scores, respectively), which is evidence that there is selection on unobservables. Finally, we run the same regressions including the retention policy variables (i.e., the exclusion restrictions), and we find that they are not significant (p-values of .082, .543, and .158, for reading, math, and general knowledge test scores, respectively). This is helpful evidence in support of our exclusion restrictions but far from definitive given that we cannot test for their validity in the dynamic model. To make sure the assumptions on the exclusions are not driving our results, we also estimate the full model without the exclusion restrictions, and we get similar results.

IV. Identification

The primary challenge in identifying the treatment effect of grade retention in the static framework is that individuals differ in unobservable ways that help determine both selection into retention and the effect of retention. For instance, lower-ability students are more likely to be retained and may also learn at a slower rate than higher-ability students, leading to a different effect of grade retention. The problem is similar in our setting, with the added challenge that selection is dynamic and that treatment effects vary over time as well as by unobservable characteristics of the student.

We first illustrate the strategy we develop to control for dynamic selection using a three-period example in Section IV.A. Central to the credibility of our identification strategy is that we have adequately controlled for the unobservables that jointly determine selection into treatment and the effect of treatment. In Section IV.B, we illustrate how identification works in the case of a unidimensional unobservable factor that affects both selection into treatment and the outcome of treatment, and we then expand this to allow for correlated shocks to outcomes over time. This is important for capturing shocks, such as divorce, which might affect both selection into retention and the treatment effect of retention. Section IV.C then expands the argument to multidimensional abilities. This appears crucial in our setting, where students may be retained for behavioral or cognitive reasons.

A. Factor Structure

Consider a three-period example, where treatment can be taken in either of the first two periods ($R = 1, 2$); that is, students can be retained in kindergarten or first grade. The policy is evaluated according to its effect on some outcome measured at period t: $Y_i(t, r)$, for example, third-period test scores. For example, potential outcomes in period 3 can be given by

$$Y_i(3, r) = \Phi(3, r) + \epsilon_i(3, r) \text{ for } r = 1, 2, \infty,$$

and the observed outcome can be written as a function of potential outcomes and treatment indicators as

$$Y_i(3) = \Phi(3,\infty) + D_i(1)[\Phi(3,1) - \Phi(3,\infty)] + D_i(2)[\Phi(3,2) - \Phi(3,\infty)] + \epsilon_i(3,\infty) + D_i(1)[\epsilon_i(3,1) - \epsilon_i(3,\infty)] + D_i(2)[\epsilon_i(3,2) - \epsilon_i(3,\infty)]. \quad (4)$$

Equation (4) is a regression model with dummy indicators for the time at which an individual is retained. It is different from a standard binary treatment model both because there is more than one treatment indicator and because the effect of treatment is potentially heterogeneous due to the interaction of the treatment indicator with the individual-specific unobservable gains from treatment. If the decision of when to receive treatment is correlated with the unobservable (to the econometrician) gains of choosing each treatment, we have a situation with *essential heterogeneity*, in the language of Heckman, Urzua, and Vytlacil (2006). That is, essential heterogeneity exists if the students who are retained are more likely to experience higher (lower) gains from retention. Formally, in our case treatment status $D_i(r)$ and/or $D_i(r')$ may be correlated with $\epsilon_i(3,r) - \epsilon_i(3,r')$ for $r \neq r'$.

One way to account for essential heterogeneity is to recover the joint distribution of the unobservables in the selection and outcome equations, (U_i, ϵ_i). This way we can describe how the treatment effect varies across unobservable individual types. Imposing a factor structure simplifies the problem and permits us to recover the joint distribution of the unobservables. In particular, we put some structure on the residuals that determine student outcomes and selection into being retained. We assume that they can be decomposed into a component that determines both selection into treatment and outcome of treatment (we refer to these as factors θ_i and interpret them as student abilities) and random component, that is:

ASSUMPTION 2. *(Factor structure)* $\epsilon_i(t,r) = \theta_i\alpha(t,r) + \varepsilon_i(t)$ and $U_i(r) = \theta_i\rho(r) + v_i(r)$, where θ_i is a vector of mutually independent "factors," and we assume that $\varepsilon_i(t) \perp\!\!\!\perp \varepsilon_i(t')$ for all $t \neq t'$, $v_i(r) \perp\!\!\!\perp v_i(r')$ for all $r \neq r'$, and $v_i(r) \perp\!\!\!\perp \varepsilon_i(t)$ for all r and t, where $\perp\!\!\!\perp$ denotes statistical independence.[9]

If assumption 1 holds, $\alpha(t,r) = \alpha(t,\infty) = \alpha(t)$ for $r \geq t$. The factor structure assumption is a convenient dimension reduction technique: it reduces

[9] We impose assumption 2 for convenience, even though it is stronger than required. Following the analysis of measurement error models in Schennach (2004) and Hu and Schennach (2008), we can relax the strong statistical independence assumptions and replace them with a combination of general dependence and weaker mean independence assumptions. Furthermore, the assumption that $v_i(r) \perp\!\!\!\perp \varepsilon_i(t)$ for all r and t would not hold in the case in which the decision to retain is based on the same test scores used in the analysis. Our identification argument can be readily adapted to account for this case (see appendix B; appendices A and B are available online).

the problem of recovering the entire joint distribution of (U_i, ϵ_i) to that of recovering the factor "loadings" $\alpha(t, r)$ and $\rho(r)$ and the marginal distributions of the elements of θ_i and of $\varepsilon_i(t), v_i(r) \; \forall \; t, r$.

The factor structure also has an appealing interpretation, since we can now talk about a low dimensional set of common "causes."[10] The same set of unobservables that determines the effect of grade retention also determines whether a student is retained. We can then see how selection into retention and the treatment effect of retention vary by student abilities and the timing of retention.

To understand how the factor structure assumption helps address the identification problem associated with unobserved heterogeneity, consider our three-period example. If assumption 2 holds, the choice process is determined by

$$V_i(r) = \lambda(r) + \theta_i \rho(r) + v_i(r) \, .$$

We find it plausible to assume that selection into retention depends on abilities that are unobservable to the researcher because parents and teachers would know more about the ability of the student. However, as discussed in Section II, the decision rule is consistent with parents and teachers basing retention only on observable test scores. The observed outcome vectors are

$$\begin{aligned} Y_i(1) &= \Phi(1) + \varepsilon_i(1) + \theta_i \alpha(1), \\ Y_i(2) &= \Phi(2, \infty) + D_i(1)[\Phi(2, 1) - \Phi(2, \infty)] + \varepsilon_i(2) + \theta_i \alpha(2, \infty) \\ &\quad + D_i(1)\theta_i[\alpha(2, 1) - \alpha(2, \infty)], \end{aligned}$$

and

$$\begin{aligned} Y_i(3) = \; & \Phi(3, \infty) + D_i(1)[\Phi(3, 1) - \Phi(3, \infty)] + D_i(2)[\Phi(3, 2) - \Phi(3, \infty)] \\ & + \varepsilon_i(3) + \theta_i \alpha(3, \infty) + D_i(1)\theta_i[\alpha(3, 1) - \alpha(3, \infty)] \\ & + D_i(2)\theta_i[\alpha(3, 2) - \alpha(3, \infty)]. \end{aligned}$$

By the third year, students could have been retained in the first period or the second period. Essential heterogeneity is present when $\alpha(t, r) \neq \alpha(t, \infty)$. In this case, unobserved gains in the test score, $\epsilon_i(t, r) - \epsilon_i(t, \infty) = \theta_i[\alpha(t, r) - \alpha(t, \infty)]$ are individual-specific and vary based on the student's unobservable abilities. Furthermore, they also determine whether a student is retained, as $D_i(r)$ is a function of the same θ_i.

If we could recover (or condition on) the unobserved θ_i, then $D_i(1)$ and $D_i(2)$ would no longer be endogenous, and we could obtain consistent esti-

[10] See Jöreskog and Goldberger (1975) for a discussion and Carneiro et al. (2003) and Cunha, Heckman, and Navarro (2005) for recent developments.

mates of the treatment effect. This is the key intuition behind the factor model, to condition not only on observable covariates but also on the unobservable vector θ_i in order to recover the conditional independence assumption of quasi-experimental methods. In this way, the factor structure is an alternative form of matching, where the idea is to "match" based not only on variables observable to the econometrician but also on the unobservable factors.

The factor model can also be understood as a generalization of the fixed effects model. To see this, take differences between the period 2 and period 1 jth outcomes as one would do in the standard fixed effects model to difference out the individual effect θ_i. In our model, differencing is not enough; instead we get

$$Y_i(2) - Y_i(1) = \Phi(2,\infty) - \Phi(1) + D_i(1)[\Phi(2,1) - \Phi(2,\infty)] + \varepsilon_i(2) - \varepsilon_i(1) + \theta_i[\alpha(2,\infty) - \alpha(1)] + D_i(1)\theta_i[\alpha(2,1) - \alpha(2,\infty)].$$

For the differencing strategy to work, we need to impose two restrictions. First, we need to rule out essential heterogeneity, that is, $\alpha(2,1) = \alpha(2,\infty) = \alpha(2)$. This eliminates any heterogeneity in the effects of retention. Second, we need to assume that the marginal effect of θ_i does not change over time; that is, $\alpha(2) = \alpha(1) \equiv \alpha$. First-differencing eliminates θ_i only when these two restrictions hold. As more periods pass, more assumptions are required for the fixed effects model to work. For instance, to identify the effect on period 3 outcomes, we would need to impose the additional assumption that $\alpha(3,2) = \alpha(3,1) = \alpha(3,\infty) = \alpha(3)$.[11]

These assumptions that permit identification of the treatment effect using the differencing strategy also eliminate important heterogeneity. For instance, they impose that the average treatment effect is the same as the treatment on the treated, which is particularly unlikely in the case of grade retention. Thus, the factor structure approach provides an important generalization of the fixed effects approach by allowing for essential heterogeneity and that the marginal effects of abilities vary by treatment status. We further generalize from the fixed effects approach by permitting multidimensional abilities, so that retention decisions and the outcome of retention can depend both on cognitive and behavioral abilities of the students, as we discuss this further in Section IV.C. These appealing properties of the factor structure are part

[11] Alternatively, by relaxing the fixed effects assumption slightly, we could employ a double-differencing strategy. We continue to rule out essential heterogeneity, but now allow for time trends. In other words, we substitute the assumption of a time-invariant marginal effect of θ_i with $\alpha(t) = \alpha_0 + \alpha_1 t$. Under these assumptions, subtracting $Y_i(2) - Y_i(1)$ from $Y_i(3) - Y_i(2)$ would recover $\Phi(3,2) - \Phi(3,\infty)$ and $\Phi(3,1) - \Phi(3,\infty) - 2(\Phi(2,1) - \Phi(2,\infty))$. However, even with these strong assumptions, we could not separate the effect of being treated in period 1 on outcomes in periods 2 and 3.

of the reason it is increasingly used in economic studies for estimating treatment effects (see Carneiro et al. [2003], among others).

B. Single-Dimensional Ability Example

To illustrate how identification works with the factor structure assumption, first consider the simplest example in which only one factor (e.g., the first element of θ_i : $\theta_{i,1}$) affects the outcome and selection equations in period 1. The outcome in period 1 is then

$$Y_i(1) = \Phi(1) + \theta_{i,1}\alpha_1(1) + \varepsilon_i(1), \tag{5}$$

where the parameters do not depend on retention status because we assume that first period outcomes are free of selection. This assumption is natural in our context where we have exams for students in kindergarten, before retention decisions are made.[12] However, it is not crucial, as we could alternatively correct for selection using exclusion restrictions. It is straightforward to show that the joint distribution of $\epsilon_i(1) = \theta_{i,1}\alpha_1(1) + \varepsilon_i(1)$ and $U_i(1) = \theta_{i,1}\rho_1(1) + \upsilon_i(1)$ is nonparametrically identified (e.g., Heckman and Smith 1998). Because we can recover the joint distribution of the residuals from the outcome and selection equation, we can also form the different moments of the joint distribution. Further, we can form

$$\frac{E(\epsilon_i^2(1)U_i(1))}{E(\epsilon_i(1)U_i^2(1))} = \frac{\alpha_1^2(1)E(\theta_{i,1}^3)}{\alpha_1(1)E(\theta_{i,1}^3)} = \alpha_1(1),$$

where we have normalized $\rho_1(1) = 1$.[13]

Implicitly, we assume here that the distribution is not symmetric; that is, $E(\theta_{i1}^3) \neq 0$. While restrictive, this assumption allows for a compact identification proof. In our application, we have more than one test per student, in which case it is easy to relax the nonsymmetry assumption, as we show in appendix B (appendices A and B are available online), but the proof becomes more cumbersome.

With $\alpha_1(1)$ in hand, it follows from a theorem of Kotlarski (1967) that the distribution of $\theta_{i,1}$ (and of $\varepsilon_i(1)$ and $\nu_i(1)$) is nonparametrically identified.[14]

[12] The main objection to this assumption is that the sample may contain students who have already been retained. We drop these students from the sample (about 1%), as indicated in Sec. III, though our results are not sensitive to their inclusion.

[13] Given that θ_1 is latent, this normalization implies no restriction since $\theta_{i,1}\rho_1(1) = \theta_{i,1}\kappa[\rho_1(1)/\kappa]$ for any constant κ.

[14] The theorem states that, if X_1, X_2, and X_3 are independent real-valued random variables, and we define

$$Z_1 = X_1 - X_2,$$
$$Z_2 = X_1 - X_3,$$

For example, suppose these distributions are such that they can be characterized by their moments (see Billingsley [1995] for conditions). Then, intuitively, identification of the distribution of $\theta_{i,1}$ follows from the fact that we can recover all its moments from taking higher-order cross moments of the residual from the selection and outcome equations; that is, $E(\epsilon_i^k(1)U_i(1)) = \alpha_1^k(1)E(\theta_{i,1}^{k+1})$ for $k > 0$.[15]

Because we are considering a dynamic model, an additional concern for identification is unobserved correlated shocks, such as parental divorce or loss of job, which might affect both selection into retention and the outcome of retention. As a consequence, we now allow for a new element of θ_i ($\theta_{i,2}$) to enter the model. The second-period outcomes and the selection equation for next period also depend on $\theta_{i,2}$ and are given by

$$Y_i(2, r) = \Phi(2, r) + \theta_{i,1}\alpha_1(2, r) + \theta_{i,2}\alpha_2(2, r) + \varepsilon_i(2), \text{ for } r \in \{1, \infty\},$$
$$V_i(2) = \lambda(2) + \theta_{i,1}\rho_1(2) + \theta_{i,2}\rho_2(2) + \upsilon_i(2).$$

Thus the $\theta_{i,2}$ is an unobserved correlated shock that affects outcomes and selection equations from period 2 onward, with the potential that its effect may change as time elapses; that is, the effect of parental divorce may be stronger in the short run than in the long run.

First, consider the selection correction for $Y_i(2, r)$. Since the retention decision for period 2 is made in the previous period, the only source of selection in this equation is $\theta_{i,1}$. Controlling for selection involves finding the distribution of $\theta_{i,1}$ conditional on the previous retention decision, as we just did above. Hence, we can control for selection even though $\theta_{i,2}$ is present. Next, we can identify the elements associated with $\theta_{i,1}$ in the period 2 equations by taking cross moments over time (i.e., $Y_i(1)$ with the selection corrected $Y_i(2, r)$ and $V_i(2)$). Finally, we can identify the elements associated with the correlated shock ($\theta_{i,2}$), as well as the nonparametric distributions of $\theta_i(2), \nu_i(2), \varepsilon_i(2)$ by taking cross moments within period 2 equations. Further periods follow similarly, with additional shocks introduced each period.

C. Multidimensional Abilities

We extend this analysis to the case in which unobserved ability ($\theta_{i,1}$) is multidimensional beyond the correlated shocks described above. We consider a normalization of θ_i such that true ability at the initial period consists of three independent components (A_i, B_i, C_i). In particular, assume we have

---

then, if the characteristic function of (Z_1, Z_2) does not vanish, the joint distribution of (Z_1, Z_2) determines the distributions of X_1, X_2, and X_3 up to location. For a proof, see Kotlarski (1967) or Prakasa Rao (1992), theorem 2.1.1.

[15] Formally, one wants to characterize a distribution using its characteristic function and not moments, and this is precisely what the Kotlarski argument does.

access to $N_c \geq 2$ measures (or tests) of cognitive functions $\zeta_{i,j}$, and $N_b \geq 2$ measures of behavioral functions, $\beta_{i,j}$, that are measured free of selection. As before, we keep all conditioning on covariates implicit to simplify notation. We write the jth demeaned period 1 cognitive test as

$$\zeta_{i,j,1} = A_i \alpha_{\zeta,j,1} + C_i \pi_{\zeta,j,1} + \varepsilon_{i,\zeta,j,1}, \tag{6}$$

and the jth demeaned behavioral test as

$$\beta_{i,j,1} = A_i \alpha_{\beta,j,1} + B_i \phi_{\beta,j,1} + \varepsilon_{i,\beta,j,1}. \tag{7}$$

The latent index for selection for period 1 is

$$V_i(1) = \lambda_{0,1} + Z_{i,1}\lambda_{z,1} + A_i \rho_{A,1} + B_i \rho_{B,1} + C_i \rho_{C,1} + \upsilon_{i,1},$$

where Z_i denotes an exclusion that affects the retention decision but not outcomes directly. We take science, math, and reading test scores as markers of cognitive ability C_i and general ability A_i (i.e., ζ) and the SRS ratings on students behavioral and social skills as our noisy measures of the behavioral ability B_i and general ability A_i (i.e., β). This is not to say that cognitive ability plays no role in behavioral aspects or vice versa but rather that whatever is common between these functions is captured by the general ability component A_i. The cognitive ability component C_i and the behavioral component B_i measure the part of ability that is used exclusively for the corresponding function.[16]

Semiparametric identification follows similarly to the one-factor model. We prove semiparametric identification of the model formally in appendix B. Intuitively, we now take moments across cognitive and behavioral equations to recover the period 1 α parameters and the nonparametric distribution of general ability, A_i. After removing general ability, we can then take cross-moments within cognitive tests to recover the period 1 π parameters and the distributions of cognitive ability, C_i, and the remaining residual, $\varepsilon_{i,\zeta}$. After removing general ability, we take cross moments within behavioral tests to recover the period 1 ϕ parameters as well as the nonparametric distributions of behavioral ability, B_i, and the remaining residual, $\varepsilon_{i,\beta}$. In essence A_i represents everything that correlates behavioral and cognitive scores, while B_i and C_i capture the residual correlation in behavioral and cognitive scores respectively after accounting for A_i. The residual variance in scores not captured by (A_i, B_i, C_i) is captured by ε_i.

Once we have recovered the distribution of (A_i, B_i, C_i), we can proceed to the next period. Now some students will be treated (i.e., will repeat kindergarten), and so the test scores in period 2 will be contaminated with selec-

[16] Other normalizations are possible, but the present normalization may also be applicable to other settings with multidimensional unobservables. See Bonhomme and Robin (2010) and Cunha et al. (2010) for examples.

tion. Since the only source of selection for period 2 test scores is given by (A_i, B_i, C_i), we can control for selection in period 2 test scores by controlling for how the (now-known) distribution of (A_i, B_i, C_i) for potential retainees varies due to past retention decisions. We can then repeat the arguments above and recover the period 2 loadings and the distribution of the period 2 ε's from the selection-corrected period 2 outcomes. However, since we now know the distribution of abilities in advance, we can let all three types of ability enter all equations (whether behavioral or cognitive) without having to normalize some loadings to zero. That is, the normalization that B_i only enters β-equations and C_i only enters ζ-equations need only apply to the first period.

Proceeding iteratively with the arguments above, we can recover all of the parameters and distributions in the outcomes of interest for each period. Furthermore, as in the single-dimensional ability example above, we can add elements to θ over time to allow for persistent unobserved (to the econometrician) shocks every period. By adding a new element to θ every period, we can capture any residual correlation in outcomes not captured by (A_i, B_i, C_i) and time varying loadings.

Finally, we correct for potential biases due to selective sample attrition (e.g., students moving to a different school if they know they will be retained in their current school) by adding an equation for missing data (i.e., a binary model for attrition) that depends on the same common vector θ_i.

V. Estimation

Let $\zeta_{i,j,1}$ be our jth cognitive measure for individual i in period 1 (kindergarten) and similarly for behavioral measures. Let $X_{i,t}$ denote the set of observable covariates related to the student, his family, his school, and his class, as described in table 1. These covariates are assumed to be orthogonal to $(A_i, B_i, \text{and } C_i)$. Our kindergarten measures are modeled as:[17]

$$\zeta_{i,j,1} = X_{i,1}\gamma_{\zeta,j,1} + A_i\alpha_{\zeta,j,1} + C_i\pi_{\zeta,j,1} + \varepsilon_{i,\zeta,j,1}, \tag{8}$$

and

$$\beta_{i,j,1} = X_{i,1}\gamma_{\beta,j,1} + A_i\alpha_{\beta,j,1} + B_i\phi_{\beta,j,1} + \varepsilon_{i,\beta,j,1}. \tag{9}$$

[17] We follow the identification arguments in Sec. IV.A and, without loss of generality, impose the following normalizations. We normalize the general ability loading on the first period general knowledge test to 1, so A can be interpreted as a trait that is associated positively with higher scores in the general knowledge test. The loading on cognitive ability is normalized to 1 on the first period math test, so C is associated with higher math scores. Finally, we normalize the behavioral loading on the self-control marker to 1.

Observed test scores in the following years are

$$\begin{aligned}\zeta_{i,j,t} = {} & X_{i,t}\gamma_{\zeta,j,t} + A_i\alpha_{\zeta,j,\infty,t} + B_i\phi_{\zeta,j,\infty,t} + C_i\pi_{\zeta,j,\infty,t} + \sum_{\tau=2}^{t}\eta_i^{(\tau)}\delta_{\zeta,j,t}^{(\tau)} + \varepsilon_{i,\zeta,j,t} \\ & + \sum_{r=1}^{t-1}D_i(r)\left[\Phi_{t,r} + A_i\left[\alpha_{\zeta,j,r,t} - \alpha_{\zeta,j,\infty,t}\right] + B_i\left[\phi_{\zeta,j,r,t} - \phi_{\zeta,j,\infty,t}\right]\right. \\ & \left. + C_i\left[\pi_{\zeta,j,r,t} - \pi_{\zeta,j,\infty,t}\right]\right],\end{aligned} \tag{10}$$

We restrict the observable covariates (except for the constant) to have the same marginal effect across time for a given subject. The main reason we do this is to save on the number of parameters we are estimating. Furthermore, preliminary reduced form regressions suggested that the marginal effects did not vary much across grades. We also restrict the effect of the permanent shock ($\eta_i^{(\tau)}$) to be the same regardless of retention status. Parameter $\Phi_{t,r}$ then measures the average effect of being retained at r in period t. Importantly, note that this specification corresponds to the general case discussed above in that the treatment varies over time as does the effect of the unobservable "abilities" (i.e., the difference in the loadings). Hence, the effect of treatment is both heterogeneous and time-varying.

To model selection into retention,[18] we write the latent index V as

$$\begin{aligned}V_i(r) = {} & \lambda_{0,r} + X_{i,r}\lambda_{x,r} + Z_{i,r}\lambda_{z,r} + A_i\rho_{A,r} + B_i\rho_{B,r} + C_i\rho_{C,r} + \sum_{\tau=2}^{r}\eta_i^{(\tau)}\psi_r^{(\tau)} \\ & + v_{i,r} \text{ for } r = 1, \ldots, \overline{R}.\end{aligned}$$

We then define $D_i(R_i)$ as

$$D_i(R_i) = \mathbf{1}\left(V_i(R_i) > 0 \mid \{V_i(r) \le 0\}_{r=1}^{R_i-1}\right).$$

Notice that we allow for exclusions in the index, so that some variables (Z) are included in the retention equations but not in the outcomes. This corresponds to the seven binary measures of the retention policies discussed in Section III.

As discussed in Section IV.A, given that test scores in kindergarten are free of selection, the additional assumption of valid exclusion restrictions is not necessary but rather aids in identification. Similarly, given valid exclusions, the assumption of initial test scores free of selection is not necessary for identification. We estimate the model with and without exclusions and find similar results, though the loglikelihood ratio test supports that the ex-

[18] Since we know the latent index is nonparametrically identified, we could instead write it as a polynomial on the variables instead of a linear function for example. Given that the number of parameters we are estimating is already 616, and the number of parameters would increase considerably, we maintain the linearity assumption.

clusions are important. Finally, to address nonrandom sample attrition based on unobservables, we also include a similar selection equation for students who select out of the sample.

The distributions of the unobservables $(A, B, C, \{\eta^{(\tau)}\}_{\tau=2}^{\bar{t}}, \varepsilon, \upsilon)$ in the model are nonparametrically identified, as shown in Section IV.A. However, for estimation purposes, we specify all of the distributions and allow them to follow mixtures of normals with either two or three components. Furthermore, while our identification arguments are presented in a sequential fashion and lead naturally to a multistep estimation procedure, we estimate all of the parameters in the model jointly by maximum likelihood in a single step primarily for efficiency reasons.[19]

VI. Results

We find that our model fits the means and variances of all the test measures very well; we cannot reject that the values predicted by the model equal those in the data. The same is true for the probabilities of retention in the data.[20]

Figure 1 presents evidence of selection on the different components of ability. Ignoring kindergartners for the moment, we find that early retainees have lower ability (by all measures) than later retainees who have lower ability than students who are not retained. This is consistent with a dynamic selection model in which you first retain the lowest-ability students and then in the next round the next-lowest ability, and so forth. Kindergarten retention appears to be an exception in that these students have higher general and behavioral ability than early retainees but lower ability than late retainees in all dimensions. This seems to suggest that the decision to retain students in kindergarten is different than in other grades. This evidence provides important support for our method, that is, the need to account for both dynamic selection and multidimensional abilities.

Table 4 describes one parameter of interest—the treatment on the treated (and the untreated) parameters for both reading and math test scores (panels A and B, respectively) in the last year in our data, the 2003–4 school year, when students are approximately age 11.[21] The columns correspond to

[19] Maximization of the log likelihood was performed in three steps. First, we use a simulated annealing algorithm, which is a probabilistic search algorithm that allows us to rule out regions of lower-valued local optima. Second, we employ a Nelder-Mead polytope search algorithm to make quick strides toward the optimum avoiding the costly calculation of gradients. Finally, we use the BFGS (Broyden–Fletcher–Goldfarb–Shanno) method to obtain convergence. We also tried several different initial values, and found the same maximum.

[20] In tables A1 and A2 we present evidence of the fit of the model. Parameter estimates and standard errors are available in online appendix A.

[21] The predicted levels of achievement from which these gains are calculated are shown in table A5.

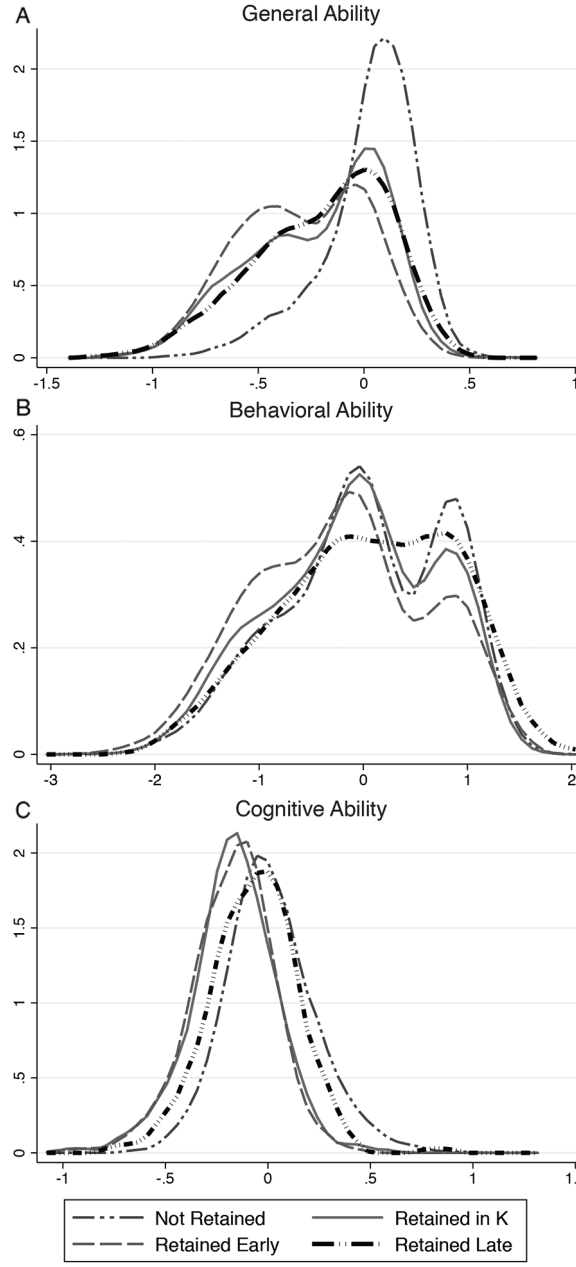

FIG. 1.—Densities of abilities by retention status. NOTE: Let $f(X)$ denote the probability density function of ability $X \in \{A, B, C\}$. We allow $f(X)$ to follow a mixture of normals distribution. Let $R \in \{1, 2, 3, \infty\}$ denote retention states: retained in kindergarten, retained early (grade 1 or grade 2), retained late (grade 3 or grade 4), and not retained. The graph shows $f(X \mid R = r)$ for each retention status. A color version of this figure is available online.

Table 4
Average Test Score Gain by Retention Status: 2003–4 School Year

	Average Gain of a Student Who Is Actually Not Retained/Retained, Conditional on the Retention Status Being:				
	Not Retained	Retained in Kindergarten	Retained Early	Retained Late	ATE (Unconditional)
A. Reading score:					
Retained in kindergarten versus not retained	.034	−.057	−.086	−.023	.025
	(.014)	(.013)	(.018)	(.027)	(.012)
Retained early versus not retained	.058	−.092	−.111	−.046	.046
	(.019)	(.019)	(.023)	(.046)	(.017)
Retained late versus not retained	.058	.026	.016	.022	.056
	(.112)	(.058)	(.080)	(.084)	(.101)
B. Math score:					
Retained in kindergarten versus not retained	.011	−.057	−.084	−.071	.004
	(.024)	(.019)	(.021)	(.031)	(.022)
Retained early versus not retained	.079	−.058	−.095	−.016	.066
	(.021)	(.015)	(.017)	(.036)	(.019)
Retained late versus not retained	.098	−.075	−.112	−.052	.083
	(.337)	(.142)	(.162)	(.258)	(.309)

NOTE.—Let $R = 1, 2, 3$, or ∞ represent the actual retention status of a student, respectively, retained in kindergarten, retained early (at grade 1 or grade 2), retained late (at grade 3 or grade 4), or never retained. Let $\zeta(i)$ be the potential test score if the student were retained at time $i = 1, 2, 3, \infty$. The row i, column j, element of this table calculates $E[\zeta(i) - \zeta(\infty) \mid R = j]$. For example, the math test score of a student who was actually not retained would increase by 0.079 if he/she were retained at grade 1 or grade 2 instead. Standard errors are in parentheses.

actual treatment statuses, whereas the rows compare potential gains across treatment statuses relative to not being retained. In other words, the first row describes the treatment effect of being retained in kindergarten versus not being retained. The last column describes the average treatment effects.

Considering, first, the treatment on the treated parameters, students who are actually retained in kindergarten perform 6% lower in reading and math by 2003–4 than if they had not been retained. This does not mean that students who are retained lose acquired knowledge by being retained, but rather that by age 11 (i.e., in 2003–4) a pair of identical students (one of whom was retained) would both have higher test scores than they did at age 6. The retained student's age 11 score, however, would be 6% lower than his counterpart. Students who are retained early perform about 11% lower in reading and 10% lower in math than if they had not been retained. The results for late retention vary across math and reading, with late retainees experiencing gains of 2% in reading but losses of 5% in math, although these

results are not statistically significantly different from 0. The off-diagonal elements of the table provide other interesting counterfactuals. For instance, a student who is retained in kindergarten would have been even worse off in reading if retained early instead. Generally these results are supportive evidence that conditional on retaining a student, the timing was optimal, that is, the student would not have been better off if retained at another time.

Overall, the treatment on the treated parameters show a negative effect of retention. In contrast, the average treatment effects reported in the last column predict that the effect of retention in kindergarten is small or zero and positive for early retention. Again, the effect is not statistically significantly different from zero for late retention. Our results show that these nonnegative average treatment effects are driven by the untreated students, for whom the treatment effect of retention is generally positive. Below we provide some intuition behind this finding, but first we consider another important piece of this puzzle, how treatment effects vary by unobserved ability.

A. Heterogeneity in Treatment Effects by Abilities

Figures 2, 3, and 4 break out the average treatment effects from table 4 to show how the treatment effects of being retained at different grades vary across the percentiles of the general, behavioral, and cognitive ability distributions for reading and math for the 2003–4 academic year. Comparing across graphs, we see that lower-ability students generally experience losses (or are no better off) due to retention, whereas the higher-ability students actually benefit from retention.

There could be several reasons for these findings. We begin by ruling out explanations related to misspecification and limited support in our data. First, it may not be possible to estimate the effect for high-ability students if we do not observe high-ability students being retained in our data, in which case our result would be purely due to functional-form extrapolation. However, the test scores reported in the ECLS-K are not actually used to determine retention decisions. While we recognize a student as high ability from the factor decomposition of the history of his performance on these tests, his performance in the classroom could suggest otherwise. In fact, some of our retainees have above-median measured achievement. Furthermore, we test that the results for high-ability students are not just noise; confidence intervals show that the effects are often statistically significantly different from 0.

A second potential reason is that our model is restricted to be linear in ability. It could be that in reality the students close to the margin benefit, while high- and low-ability students experience losses from retention. We estimate a more flexible version of our model that includes a quadratic in ability in the outcome equations, thus permitting this sort of inverted-U-shaped pattern in ability. While we do find evidence of some inverted U's,

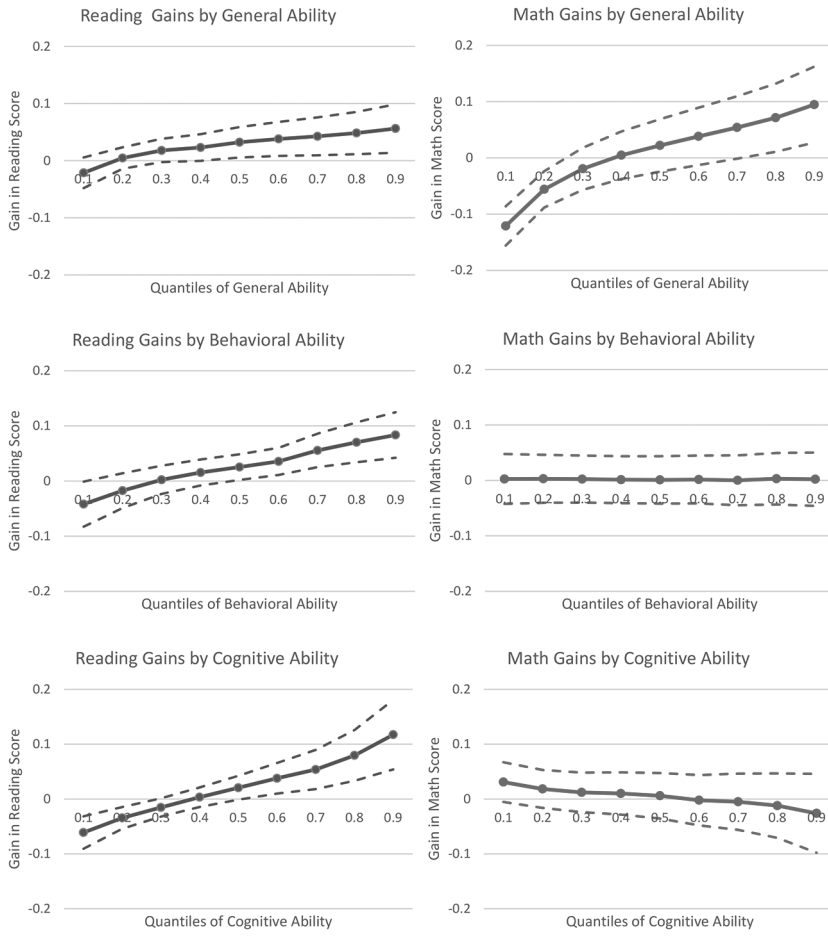

FIG. 2.—Achievement gains (retained in kindergarten vs. not retained) in 2003–4 by ability quantiles. NOTE. Let $\zeta(t, 1)$ and $\zeta(t, \infty)$ be the potential test scores at period t if the student is retained in kindergarten and if the student is not retained at all, respectively. Let X denote one kind of ability: $X \in \{A, B, C\}$. The graph shows $E(\zeta(t, 1) - \zeta(t, \infty) \mid X = q)$, where q is the qth quantile of the X –type of ability distribution. Dashed lines show the 95% confidence interval. A color version of this figure is available online.

this is far from being a consistent pattern. In some cases, the upward-sloping treatment effects in ability become even more pronounced. Furthermore, model selection tests favor the linear model over the quadratic ones.

Another reason could be that higher-ability students actually benefit more from retention than low-ability students. We find that the factor loadings are larger for the retained than for the not retained outcomes and pos-

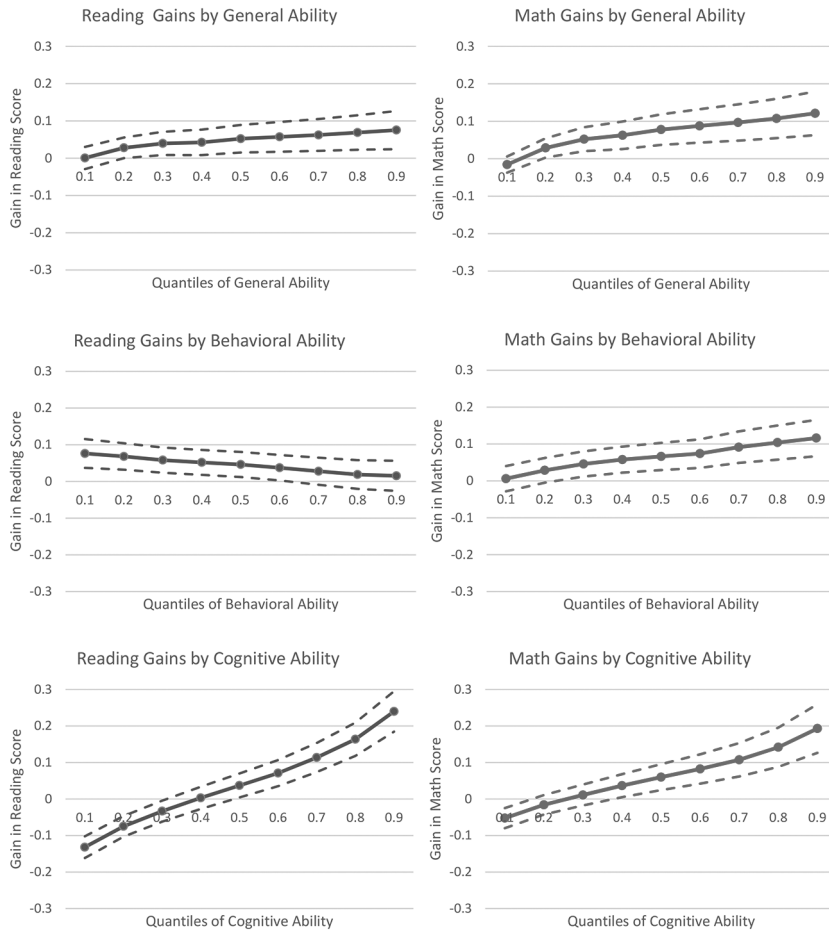

FIG. 3.—Achievement gains (retained early vs. not retained) in 2003–4 by ability quantiles. NOTE. Let $\zeta(t, 2)$ and $\zeta(t, \infty)$ be the potential test scores at period t if the student is retained early (grade 1 or grade 2) and if the student is not retained at all, respectively. Let X denote one kind of ability: $X \in \{A, B, C\}$. The graph shows $E(\zeta(t, 2) - \zeta(t, \infty) \mid X = q)$, where q is the qth quantile of the X –type of ability distribution. Dashed lines show the 95% confidence interval. A color version of this figure is available online.

itive in cognitive and general ability (see table A10; tables A1–A19 are available online). Given that ability has mean 0, this means roughly that high-ability students experience achievement gains relative to not being retained, whereas low-ability students experience losses relative to not being retained. There are several economics-based explanations for this that are supported both in our data and in the literature. We present empirical evidence for the

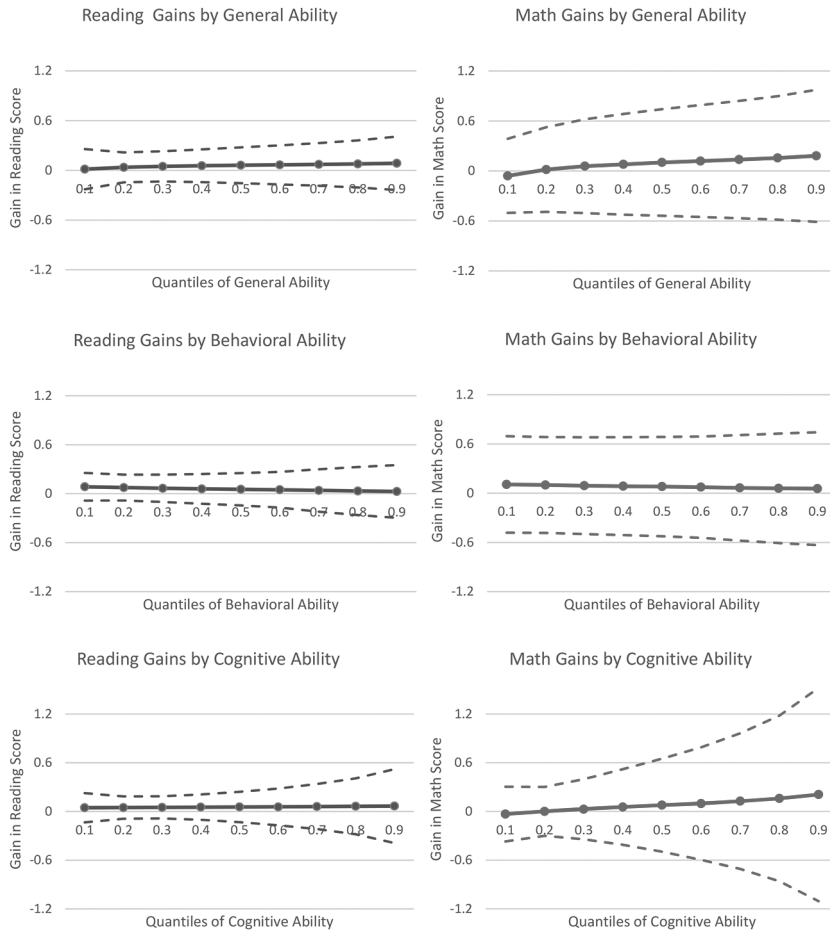

FIG. 4.—Achievement gains (retained late vs. not retained) in 2003–4 by ability quantiles. NOTE. Let $\zeta(t, 3)$ and $\zeta(t, \infty)$ be the potential test scores at period t if the student is retained late (grade 3 or grade 4) and if the student is not retained at all, respectively. Let X denote one kind of ability: $X \in \{A, B, C\}$. The graph shows $E(\zeta(t, 3) - \zeta(t, \infty) \mid X = q)$, where q is the qth quantile of the X –type of ability distribution. Dashed lines show the 95% confidence interval. A color version of this figure is available online.

most intuitively appealing ones: that high-ability students who are retained either receive more investment or can better take advantage of opportunities post-retention than their low-ability counterparts.

First, high-ability students may have higher-ability parents (assuming intergenerational transmission of human capital). We find evidence of this in our data; higher-ability students who are retained in kindergarten come

from higher-SES families and are less likely to be from single-parent families.[22] Higher-SES parents may be better equipped to ensure that when their child is retained he/she gets the best teachers and the resources he needs. Thus, resources may be invested disproportionately more in high-ability students who are retained than in low-ability students. Note that this interpretation is in no way inconsistent with the assumption that unobservable abilities of students are orthogonal to observable covariates; that is, abilities conditional on retention status may be correlated with observable covariates. We test for this directly using a difference-in-difference strategy and find some evidence to support this hypothesis. Higher-ability students who are retained in kindergarten experience larger increases in the quantity of books in the home in the next year and are more likely to have a TV rule put in place, relative to lower-ability students who are retained. To the extent that race proxies for SES, this is also supported by Eide and Showalter (2001), which finds positive effects of grade retention for whites but not nonwhites, and Jacob and Lefgren (2009), which finds more negative effects primarily for black females in eighth grade.

Furthermore, on average, high-ability students may attend better schools and/or have more resources at their disposal than low-ability students who are retained, further reinforcing our argument. We find some evidence in the data that higher-ability students who are retained in kindergarten have more resources at their disposal relative to lower-ability students in the form of more books in the home and smaller class sizes.

Additionally, even if teachers and/or parents put more resources into students who are retained equally, we may still observe this pattern. If high-ability students are better equipped to take advantage of these resources than low-ability students, this may explain the difference across ability types. High-ability students also may benefit from being retained if, by being retained, they are put in the position of teaching other students or gain confidence as they see that they are able to perform well next to the new cohort of students. In contrast, low-ability students who are retained may not be in a position to offer help to their new cohort of peers. They may even lose self-esteem if they find that they continue to perform worse next to their younger cohort. This finding is supported by Bedard and Dhuey (2006) and others suggesting that the age relative to other students in the classroom matters for performance.

Importantly, while we provide support for our finding that high-ability students who are retained benefit relatively more than lower-ability students, we would not conclude from our findings that in general high-ability students should be retained, for several reasons. First, we can only esti-

[22] For each student in our sample, we use the model to predict their abilities given all the information (i.e., test scores, retention decisions, covariates) we observe by using Bayes's rule repeatedly.

mate the effect of retention on the support of students who are actually retained. While there appear to be some relatively high-ability students retained in our data, as argued above, the results may not generalize to the highest-ability students. Second, the negative consequences of the year lost by a high-ability student from retention in terms of wages and additional schooling could easily outweigh the achievement benefits we estimate in our data. Third, the model is not a general equilibrium model and clearly could not accurately predict the effect of retaining all high-ability students.

On the other end, it is also important to point out that grade retention generally has a negative effect on the low-ability students in our sample. These findings in particular are a useful addition to the literature, as previous studies that have dealt with selection on unobservables, in particular regression discontinuity estimates, cannot speak to these lowest-ability students (e.g., Jacob and Lefgren 2004, 2009; Manacorda 2012), and these students are often the target of retention policies.[23]

B. Time-Varying Treatment Effects

The results so far also illustrate considerable heterogeneity in treatment effects across retention times. On the one hand, this heterogeneity would follow if there is something substantively different about retention at these different grades, such as the repetition of first grade producing larger benefits on average than the repetition of kindergarten. On the other hand, it could be that the disparities are driven by the time elapsed since retention and our choice to focus on 2003–4 outcomes. For instance, for the case of late retention, the results reported in table 4 and figures 2–4 are short-run effects, achievement gains 1–2 years after retention. For kindergarten retention, the effects are longer run, that is, 4–5 years after treatment.

To consider how treatment effects vary over time, figures 5 and 6 compare treatment effects of kindergarten and early retention at the different periods we observe in the data. The left-hand-side figure depicts the evolution over time of the average treatment effect and the right-hand-side figure depicts the treatment on the treated for kindergarten and early retention, re-

[23] Jacob and Lefgren (2004) do look at heterogeneity in effects by prior achievement (the year before the threshold passing rule was implemented). They find some evidence of negative effects for the lowest and highest achieving in math, but no statistically significant effects in reading. While these are useful estimates, one concern is that the source of variation in their performance from year to year is not random and may be driven by shocks, like divorce, parental loss of job, etc. This would suggest that these students are special cases and caution should be exercised in interpreting heterogeneity in regression discontinuity results based on prior achievement. Our method not only controls for these types of correlated shocks, but recovers a more permanent type of ability that is based on the history of performance, rather than just a single observation. In principle, these estimates could be more useful from a policy perspective in that we expect parents to know more about ability and to base retention decisions in part on the history of their child's performance.

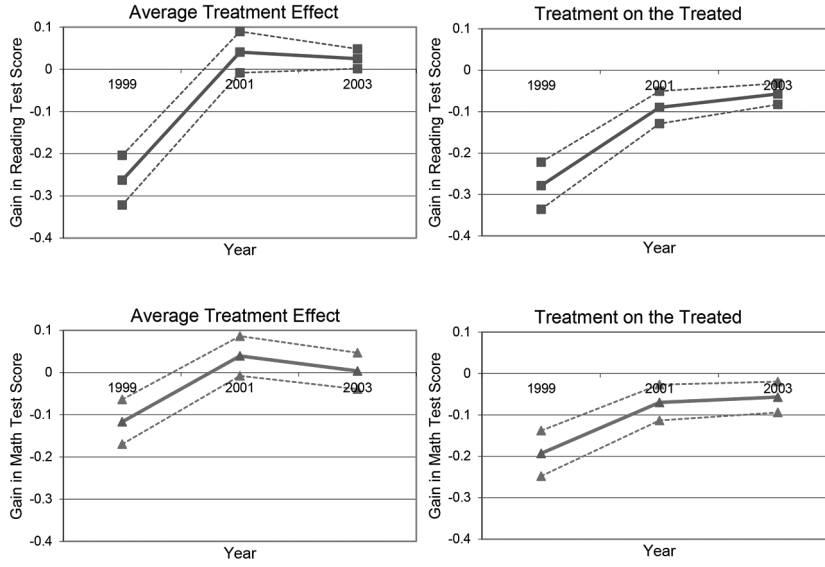

FIG. 5.—Achievement gains for kindergarten retention over time. NOTE. Let $\zeta(t, 1)$ and $\zeta(t, \infty)$ be the potential test scores at period t if the student is retained in kindergarten and if the student is not retained at all, respectively. Let $R \in \{1, 2, 3\}$ indicate the period a student is retained at. The average treatment effect graph shows $E(\zeta(t, 1) - \zeta(t, \infty))$ for $t = 1, 2$, and 3 for each test score. The treatment on the treated graph shows $E(\zeta(t, 1) - \zeta(t, \infty) \mid R = t)$. Dashed lines show the 95% confidence interval. A color version of this figure is available online.

spectively.[24] Figure 5 shows that the initial effect of being retained in kindergarten is fairly strongly negative, with students performing on average 26% lower in reading and 12% lower in math than if they had not been retained. However, 2 years later (in 2001) the average treatment effect is somewhat positive at 4%, and it goes down to 3% for reading and 0 for math in 2003. Thus, while the initial effect of retention is negative and large, students on average appear to catch up in the long run.

The right-hand-side panel of figure 5 shows a similar pattern for the treatment on the treated, that is, students who are actually retained in kindergarten. The initial effect of retention is slightly more negative than for the average student, −28% in reading and −19% in math. Like the average student, the treated students have made significant progress 2 years later and only perform about 9% lower in reading and 7% lower in math than if they had not been retained. However, the treatment on the treated remains neg-

[24] Tables A8 and A9 show the gains and standard errors for different time periods and correspond to the different points in these figures.

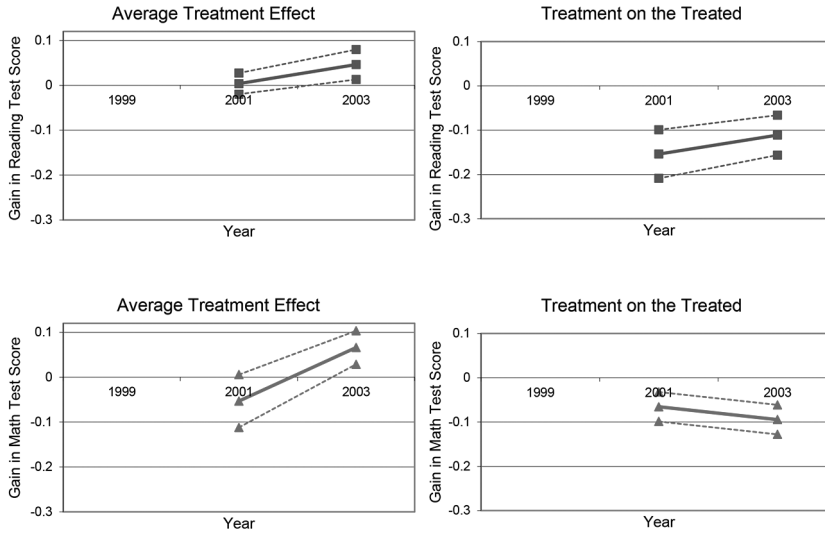

FIG. 6.—Achievement gains for early retention over time. NOTE. Let $\zeta(t, 2)$ and $\zeta(t, \infty)$ be the potential test scores at period t if the student is retained early (grade 1 or grade 2) and if the student is not retained at all, respectively. Let $R \in \{1, 2, 3\}$ indicate the period a student is retained at. The average treatment effect graph shows $E(\zeta(t, 2) - \zeta(t, \infty))$ for $t = 1$, 2, and 3 for each test score. The treatment on the treated graph shows $E(\zeta(t, 2) - \zeta(t, \infty) \mid R = t)$. Dashed lines show the 95% confidence interval. A color version of this figure is available online.

ative in 2003–4 (4 years later) at about −6%, so the effect does not become positive as is the case for the average student.

With early retention (fig. 6), we can only compare the short-run effect (in 2001) to the effect 2 years later (in 2003). In contrast to kindergarten retention, the initial effect of early retention for the average student is much smaller, approximately 0 for reading and −5% for math. The longer-run effect is positive, 5% for reading and 7% for math, on average. The initial effect of retention on early retainees is also less negative than the initial effect for kindergarten retainees, −15% and −7% for reading and math, respectively. As in kindergarten, there is evidence that students catch up with where their reading score would have been if not retained, but not in math. One reason that the initial negative effect of retention is smaller for early retainees could be because this effect may be measured up to 2 years after retention occurred (i.e., they could have been retained in first or second grade), whereas the estimated initial effect for kindergarten retainees is in the first year after retention.

The fact that the average treatment effect is, in general, less negative than the treatment on the treated over time is consistent with our findings in

Section VI.A that the effect on the treated student is more negative than for the average student.[25] In their meta-analysis, Allen et al. (2009) find that studies that estimate the effect of grade retention holding age fixed, as in our case, find initial losses to achievement that appear to go away over time. In contrast, studies that hold grade fixed find the opposite—initial gains to achievement that go away over time.

Similar to our findings, Jacob and Lefgren (2009), using their regression discontinuity design, find that the effects of grade retention on high school completion are not statistically significantly different from zero for students retained in sixth grade but are negative for students retained in eighth grade. They posit that the students in sixth grade have more time to catch up, an argument that is supported in our findings. While we are not able to look at long-run employment or drop-out effects, as in Jacob and Lefgren (2009), an advantage of our approach is that we are able to condition on the history of the student's performance, and even treatment status, thus providing a picture of long-run effects that are not confounded with later possibilities of grade retention, something that cannot be controlled for in the regression discontinuity framework or the studies in the meta-analysis. Jacob and Lefgren (2009) also point out that students who manage to just pass in sixth grade are more likely to be retained in eighth grade, potentially confounding estimates of the comparison of the treatment effect of retention across grades.

C. Comparison with Estimated ATE Using OLS and FE

To help place our estimates in context, table 5 compares average treatment effects in reading scores (panel A) and math score (panel B) using ordinary least squares, fixed effects, and our factor method. The model is estimated jointly in each case, allowing a separate effect of retention in different years. For OLS, the math scores are used to attempt to control for selection (or unobservable "ability") in the reading equation, and reading scores attempt to control for selection in the math equation. While the treatment on the treated may be the more interesting comparison, the OLS and fixed effects estimators are poorly equipped for these comparisons.

Considering reading scores in panel A, the initial effect of kindergarten retention on reading in 1999–2000 is negative and takes similar values across estimation methods, ranging from −24% with OLS, −26% with our method, and −28% using individual fixed effects. However, by 2001–2 (col. 3), the results become qualitatively different across the methods. The OLS model predicts that achievement is 7% lower for students retained in kindergarten, whereas our model predicts that it is 4% higher. The fixed

[25] Figures A1–A3 (available online) show the effects by ability at different years and show again that higher-ability students generally fare better than low-ability students when retained regardless of how much time has elapsed since retention.

Table 5
Estimated Coefficients for Retention Variables in Outcome Equation

	Outcome Equation in 1999–2000 School Year (1)	Outcome Equation in 2001–2 School Year (2)	Outcome Equation in 2003–4 School Year (3)
A. Reading score:			
Retained in kindergarten:			
OLS	−.241	−.068	−.065
	(.015)	(.019)	(.021)
Fixed effect	−.283	−.008	.051
	(.018)	(.021)	(.024)
Model	−.263	.041	.025
	(.030)	(.025)	(.012)
Retained early:			
OLS		−.146	−.080
		(.020)	(.019)
Fixed effect		−.049	.062
		(.021)	(.020)
Model		.004	.046
		(.012)	(.017)
Retained late:			
OLS			.014
			(.038)
Fixed effect			.074
			(.038)
Model			.056
			(.101)
B. Math score:			
Retained in kindergarten:			
OLS	−.025	−.050	−.049
	(.017)	(.021)	(.024)
Fixed effect	−.099	.071	.151
	(.018)	(.022)	(.024)
Model	−.117	.039	.004
	(.027)	(.024)	(.022)
Retained early:			
OLS		−.040	−.060
		(.023)	(.022)
Fixed effect		.039	.116
		(.021)	(.021)
Model		−.053	.066
		(.030)	(.019)
Retained late:			
OLS			−.091
			(.044)
Fixed effect			.075
			(.039)
Model			.083
			(.309)

NOTE.—For the OLS and fixed effect regressions to better correspond to the estimated model, they are run on the pooled data set. The coefficients for the covariates are not allowed to change over time. Year dummies and interactions of year dummies and retention indicators are included. In addition, OLS regressions control for math scores (panel A) and reading scores (panel B). Standard errors are in parentheses.

effects estimate is approximately 0. Similarly, OLS predicts a bigger negative initial effect of early retention of −15%, in contrast to smaller estimated effects of −5% for fixed effects and 0 for our model. One reason these estimates may diverge over time is because of the changing importance of different components of ability (as evidenced in the variance decomposition in tables A3 and A4). Ordinary least squares and fixed effects only control for unobservable abilities in one dimension, through contemporaneous test scores in the other subject for OLS and repeated values of tests in the same subject for the fixed effects. In contrast, the measures of ability in our model take into account the whole history of test scores, as well as control for different dimensions of ability. The fixed effects estimator also assumes that this fixed ability component affects selection in the same way over time, which we find not to be the case using our method.

By 2003–4, OLS still estimates a negative effect of kindergarten and early retention, though the negative effect of early retention is smaller in magnitude than the initial effect in 2001–2. In contrast, the fixed effects estimator predicts a positive effect of kindergarten and early retention. Our model also predicts positive effects, but they are smaller in magnitude than the fixed effects. At the very least, this comparison suggests that our findings of positive average treatment effects are not unique to our model. Even more important, however, OLS generally predicts the wrong sign of the average treatment effect, particularly in the long run, which would lead to the erroneous conclusion that the effect of grade retention for the average student is negative. In contrast, fixed effects overstates the benefit of grade retention for the average student in the long run by as much as 15% higher returns than our model.

D. Marginal Policy Change

Because there is considerable heterogeneity in treatment effects by abilities, the effect of a marginal change in retention policy will depend on the abilities of the students affected by the change. As a result, its effect could differ considerably from the effects for the average, the average treated student or the average untreated student discussed above.

We consider the effect of a marginal change in retention policies in table 6. We simulate the effects of changing the retention policy dummies in table 2 to take value 0, making it harder for all schools to retain students. In column 1, we show the gains in achievement for those students who are no longer retained as a consequence of the policy change. For comparison, column 2 shows the average counterfactual gain to not being retained for students in the original retention status (i.e., the negative of the treatment on the treated parameter in table 4), while column 3 shows the average counterfactual gain to not being retained for students who are not retained (i.e., the negative of the treatment on the untreated parameter).

Table 6
Policy Simulation Treatment Parameters: 2003–4 School Year

Retention Status, Old Policy	Retention Status, New Policy	Average Test Score If Not Retained Minus Test Score If Retained Conditional on: Changing to Not Retained (1)	Original Retention Status (2)	Not Retained (3)
		Panel A. Reading Score		
Kindergarten	Not retained	.032	.057	−.034
		(.011)	(.013)	(.014)
Early	Not retained	.060	.111	−.058
		(.026)	(.023)	(.019)
Late	Not retained	−.097	−.022	−.058
		(.090)	(.084)	(.112)
		Panel B. Math Score		
Kindergarten	Not retained	.027	.057	−.011
		(.018)	(.019)	(.024)
Early	Not retained	.067	.095	−.079
		(.027)	(.017)	(.021)
Late	Not retained	−.032	.052	−.098
		(.268)	(.258)	(.337)

NOTE.—We fix all retention policy variables in table 2 to 0 for all individuals. That is, we make it harder for students to be retained. Let R_0 denote the retention status under the old policy and let R_1 be the retention status under the new policy. Let ζ_0 denote the test score under original policy and ζ_1 denote the test score under the new policy. Column 1 reports $E(\zeta_1 - \zeta_0 \mid R_1 \neq R_0, R_1 = \infty)$, col. 2 reports $E(\zeta_1 - \zeta_0 \mid R_0)$, and col. 3 reports $E(\zeta_1 - \zeta_0 \mid R_1 = \infty)$. Notice that while some students switch to other states besides $R_1 = \infty$ as a consequence of the policy, there are very few, and the results are harder to interpret, so we focus only on the $R_1 = \infty$ subgroup. Standard errors are in parentheses.

For example, the first row in panels A and B considers the case where students are originally retained in kindergarten but are now no longer retained because of the policy change for reading and math, respectively. In column 1, we see that these marginal students gain 3% in both reading and math from the change in retention status to not being retained. In contrast, the average student who is not retained loses 3% in reading and 1% in math by not being retained relative to being retained in kindergarten (col. 3). The average student already being retained in kindergarten gains 6% in reading and in math if he/she was not retained (col. 3). Except for the case involving late retention in reading, where the estimate is very imprecise, the point estimate of the effect for the marginal student affected by the policy lies in between the average effects for students in the original and new retention statuses.

The return to the marginal student is closer to the treatment on the treated estimate than it is to the treatment on the untreated one. This makes sense given the wider range of abilities in the untreated sample. The students affected by the policy have higher abilities than the average student already retained and lower abilities than those not retained, so they are not hurt as much by retention as the average treated student.

VII. Conclusion

Overall, consistent with the preponderance of evidence in the literature, our results suggest that grade retention is not an effective policy for raising the performance of students targeted by the policy, the lowest-ability students. In fact, with the exception of late retainees, we find that students who are retained experience considerable achievement losses relative to not being retained, as large as 28% lower achievement than they would have acquired if they had not been retained. On the more positive side, we find that retained students may catch up after several years, so negative effects on their achievement do not appear to persist.

We also find that the effect of repeating a grade on test scores varies considerably by student type (or ability), with the lowest-ability students generally being hurt the most by retention. We find positive effects of grade retention on higher-ability and (relatedly) untreated students in many cases; thus, the associated average treatment effects are positive. This underscores the fact that estimates of the average treatment effect, the focus of other estimators, such as fixed effects, would not be particularly relevant for policy aimed at low-ability students. We find evidence that the positive effect of grade retention for high-ability students may be due to higher resource investment conditional on retention relative to lower-ability students. We discuss some other reasons for the positive effect for high types, and we rule out that this is pure noise due to lack of support or driven by functional form assumptions.

We also find disparities in the effect of grade retention based on the timing of retention. For instance, the initial effect of kindergarten retention is about two times as negative as the initial effects of early retention. While previous studies have also recognized the importance of exploring heterogeneity in the effect of grade retention across grades by using a repeated static treatment effect framework (see Holmes [1989] and Jimerson [2001] for an overview and Jacob and Lefgren [2004, 2009] for more recent evidence), these studies fail to control for dynamic selection. For instance, while Jacob and Lefgren (2004) can provide useful insight into how treatment effects vary across grades given the particular threshold used for retention in these different grades, the comparison confounds the ability of the students being retained (due to dynamic selection) and the grade-specific effect of retention. This is particularly important given evidence of dynamic selection, that is, that students who are retained in first grade have lower ability in several dimensions than students who are retained in kindergarten or third or fourth grade.

In contrast to existing methods, we can directly answer questions about optimal timing, in the sense of whether a student retained in first grade, for instance, would be better off being retained in kindergarten or later. We find that generally, conditional on the decision to retain a student, schools tend to be making the right decision about when to retain that student.

Interestingly, studies using regression discontinuity have found effects ranging from negative (Jacob and Lefgren [2009] and Manacorda [2012] for eighth-graders), to not statistically significantly different from 0 (Jacob and Lefgren [2004] for sixth-graders), to positive (Jacob and Lefgren [2004)] for third-graders). Our findings suggest one potential reason for this disparity is that some promotion thresholds may target higher-ability students than others, leading to different estimates of the treatment effect. These disparities could also be driven by the time elapsed since retention or the grade at retention, which our study seeks to inform.

The method we present can be applied to identify causal treatment effects in many other settings where heterogeneity in the effect of treatment across time/treatments and unobservables is likely to be important. Many policy evaluation problems involve multiple potential treatments, whether time is involved or not. These cases do not fit naturally into the standard binary treatment framework that has become the workhorse of the literature, and the analyst faces similar challenges as those highlighted in our application.

References

Abbring, Jaap H., and Gerard J. Van den Berg. 2003. The nonparametric identification of treatment effects in duration models. *Econometrica* 71, no. 5:1491–1517.

Allen, Chiharu S., Qi Chen, Victor L. Willson, and Jan N. Hughes. 2009. Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis* 31, no. 4:480–99.

Bedard, Kelly, and Elizabeth Dhuey. 2006. The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics* 121, no. 4:1437–72.

Billingsley, Patrick. 1995. *Probability and measure*. 3rd ed. New York: Wiley.

Bonhomme, Stéphane, and Jean-Marc Robin. 2009. Consistent noisy independent component analysis. *Journal of Econometrics* 149, no. 1:12–25.

———. 2010. Generalized non-parametric deconvolution with an application to earnings dynamics. *Review of Economic Studies* 77, no. 2:491–533.

Brodaty, Thomas, Robert J. Gary-Bobo, and Ana Prieto. 2008. Does speed signal ability? The impact of grade repetitions on employment and wages. CEPR Discussion Paper no. 6832, Center for Economic Policy Research, London.

Carneiro, Pedro, Karsten Hansen, and James J. Heckman. 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice: The 2001 Lawrence R. Klein lecture. *International Economic Review* 44, no. 2:361–422.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. The value of school facility investments: Evidence from a dynamic regression discontinuity design. *Quarterly Journal of Economics* 125, no. 1: 215–61.

Cunha, Flavio, James J. Heckman, and Salvador Navarro. 2005. Separating uncertainty from heterogeneity in life cycle earnings: The 2004 Hicks lecture." *Oxford Economic Papers* 57, no. 2:191–261.

———. 2007. The identification and economic content of ordered choice models with stochastic cutoffs. *International Economic Review* 48, no. 4: 1273–1309.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78, no. 3:883–931.

Dong, Yingying, and Arthur Lewbel. 2015. Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics* 97, no. 5:1081–92.

Eide, Eric R., and Mark H. Showalter. 2001. The effect of grade retention on educational and labor market outcomes. *Economics of Education Review* 20, no. 6:563–76.

Eurydice. 2011. Grade retention during compulsory education in Europe: Regulations and statistics. Thematic Reports, European Commission, Education, Audiovisual, and Culture Executive Agency.

Fertig, Michael. 2004. Shot across the bow, stigma or selection? The effect of repeating a class on educational attainment. RWI Discussion Paper no. 0019, Rheinisch-Westfälisches Institut für Wirtschaftsforschung.

Fruehwirth, Jane Cooley, and Jeffrey Traczynski. 2013. Spare the rod? School responses to repeatedly failing to meet accountability standards. Unpublished manuscript, Cambridge University.

Gary-Bobo, Robert, Marion Gousse, and Jean-Marc Robin. 2013. Grade retention and unobserved heterogeneity. Unpublished manuscript, CREST-ENSAE, Center for Research in Economics and Statistics.

Greene, Jay P., and Marcus A. Winters. 2007. Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy* 2, no. 4:319–40.

Ham, John C., and Robert J. LaLonde. 1996. The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica* 64, no. 1:175–205.

Heckman, James J., and Salvador Navarro. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* 86, no. 1:30–57.

———. 2007. Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics* 136, no. 2:341–96.

Heckman, James J., and Jeffrey A. Smith. 1998. Evaluating the welfare state. In *Econometrics and economic theory in the twentieth century: The Ragnar*

Frisch centennial symposium, ed. Steinar Strom, 241–318. New York: Cambridge University Press.
Heckman, James J., Sergio Urzua, and Edward J. Vytlacil. 2006. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88, no. 3:389–432.
Holmes, C. T. 1989. Grade-level retention effects: A meta-analysis of research studies. In *Flunking grades: Research and policies on retention*, ed. Lorrie A. Shepard and Mary Lee Smith, 16–33. London: Falmer.
Hu, Yingyao, and Susanne M. Schennach. 2008. Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76, no. 1: 195–216.
Jacob, Brian A., and Lars Lefgren. 2004. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics* 86, no. 1:226–44.
———. 2009. The effect of grade retention on high school completion. *American Economic Journal: Applied Economics* 1, no. 3:33–58.
Jimerson, Shane R. 2001. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review* 30, no. 3:420–37.
Jöreskog, Karl G., and Arthur S. Goldberger. 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 70, no. 351:631–39.
Kotlarski, Ignacy. 1967. On characterizing the gamma and normal distribution. *Pacific Journal of Mathematics* 20, no. 1:69–76.
Manacorda, Marco. 2012. The cost of grade retention. *Review of Economics and Statistics* 94, no. 2:596–606.
Navarro, Salvador. 2008. Control function. In *The new Palgrave dictionary of economics*, 2nd ed., ed. Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan.
NCES. 2009. Condition of education 2009, indicator 18. Technical report, National Center for Education Statistics, Washington, DC.
Prakasa Rao, B. L. S. 1992. *Identifiability in stochastic models: Characterization of probability distributions*. Probability and mathematical statistics. Boston: Academic Press.
Rokkanen, Miikka. 2014. Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design. Unpublished manuscript, Columbia University.
Schennach, Susanne M. 2004. Estimation of nonlinear models with measurement error. *Econometrica* 72, no. 1:33–75.
Urzua, Sergio. 2013. Heterogeneous economic returns to postsecondary degrees: Evidence from Chile. NBER Working Paper no. 18817, National Bureau of Economic Research, Cambridge, MA.