Identification and Estimation of Time-Varying Treatment Effects: How The Timing of Grade Retention Affects Outcomes[*]

Jane Cooley[†], Salvador Navarro[‡]and Yuya Takahashi[§]

March 10, 2010

Abstract

This paper builds a framework for analyzing models with multiple time-varying treatments when selection into treatment is sequential and varies across treatment statuses. The key challenge is to separate the time-varying effect of treatment from dynamic selection into treatment based on unobservables. To address this problem, we develop a method that is a hybrid between a control function and a generalized fixed effects approach. Using a factor structure, we recover the distribution of the unobservables that jointly determine selection into treatment and the effect of treatment. A useful feature of our method is that it can recover the distribution of heterogeneous treatment effects across unobservable types. We apply our strategy to study grade retention using the Early Childhood Longitudinal Study of Kindergartners. We find evidence of dynamic selection into retention and that the treatment effect of retention varies considerably across grades and unobservable abilities of students.

Keywords: time-varying treatments, dynamic selection, grade retention.

[*] First draft: December 26, 2007. Navarro's work was supported by the Institute for Research on Poverty at the University of Wisconsin-Madison. We thank Pat Bajari, Anirban Basu, Buz Brock, Juan Esteban Carranza, Steven Durlauf, Han Hong, Caroline Hoxby, David Lee, David Meltzer, Dan Millimet, Karl Scholz, Jeff Smith, Jim Walker, Ken Wolpin, participants at numerous seminars and especially John Kennan and Chris Taber for providing helpful comments.

[†] Department of Economics, University of Wisconsin-Madison. email: jcooley@ssc.wisc.edu

[‡] Department of Economics, University of Wisconsin-Madison. email: snavarro@ssc.wisc.edu

[§] Department of Economics, University of Wisconsin-Madison. email: takahashi2@wisc.edu

1 Introduction

Most policy evaluation problems in social science do not fit into the simple binary treatment framework that is the focus of much of the literature. In some cases, there are multiple potential treatments. For instance, individuals who lose their job may be assigned to several different types of welfare programs with different effects on the duration of unemployment, as in Nekipelov (2008). An important special case that applies to many settings is that of time-varying treatment effects, where the effect of treatment varies according to time of treatment. In the case of the return to an advanced degree like an MBA (e.g., Arcidiacono, Cooley, and Hussey, 2008), the return may depend on the number of years that have elapsed since bachelor's completion. In the analysis of fertility, the timing and spacing of children is important (e.g., Heckman, Hotz, and Walker, 1985; Heckman and Walker, 1990). In the analysis of health outcomes, the time between a negative health shock and treatment receipt is crucial. For example, the total cost of treatment (or the survival rate) for breast cancer may be different for women who take longer to get a mastectomy after diagnosis. In the case of grade retention, it is not only whether a child repeats a grade that may affect his test scores but also the grade in which he is retained. In all of these cases, the timing of treatment may be as important a part of the decision process as whether or not to receive treatment. Furthermore, effects may differ depending on how much time has elapsed since treatment.

In this paper, we develop a simple framework for the analysis of models with multiple treatments, particularly focusing on the analysis of time-varying treatment effects. We focus on a setting where the effect of treatment varies based on the time it is received and/or the time elapsed since receipt. The leading example in our paper, and the question we address in our application, is the effect of being held back at different grades (i.e., grade retention) on student achievement. Grade retention is a controversial education policy and the evidence is mixed.(e.g., Holmes, 1989; Jimerson, 2001; Jacob and Lefgren, 2004) Yet, many nations permit and encourage grade retention to varying degrees. For instance, in the United States, it is becoming increasingly common, with the advent of nationwide accountability policies, to tie grade promotion to performance on state standardized exams. We provide new evidence on the dynamic effects of grade retention, with the goal of informing both how the timing of retention decisions affects outcomes and how effects may vary across student types.

Like in the static binary framework, the key identification challenge is separating the effect of treatment from selection of different unobservable types into treatment. The added complication in our setting is that selection is dynamic, in the sense that the decision to be treated today depends on the decision yesterday and different types may select into treatment at different times. For example, students who just miss being retained in kindergarten may be more likely to be retained in first grade. Furthermore, students who are retained in earlier grades may differ in unobservable ways from students who are retained in later grades. In the absence of adequate controls, the time-specific treatment effect of retention cannot be separated from dynamic selection.

We propose a new approach to account for dynamic selection, which can be understood as a hybrid between the control function and a generalized 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).[1] It is a control function approach because we use information from the selection equation to help control for selection. For example, the same unobserved abilities affect both test scores and the probability of being retained.

Comparable to a fixed effects approach, which controls for a time-invariant unobservable through time differences, the factor structure we propose also assumes time-invariant, individual-specific unobservable attribute. However, unlike the fixed effects approach, our factor structure permits the unobservable to have different marginal effects on selection and the outcome equations at different points in time. For instance, ability can play a more dominant role in later relative to earlier retention decisions, and the effect of ability on achievement can vary across grades. Furthermore, unlike a fixed effect which is single-dimensional, the factor structure permits the unobservable attribute to be multidimensional (e.g., behavioral and cognitive ability) and for there to be persistent shocks to outcomes over time (e.g., the effect of a bad teacher or parental divorce).

The factor structure can also be understood as an alternative form of matching where the match is based not only on observables but also on the unobservable factors. Identification of the factor structure follows both through restrictions on the covariances between unobservables in the outcome and selection equations and potentially through exclusion restrictions (variables that affect selection into treatment but not the outcome directly). The main intuition is that, by using the multiple noisy measures of the factors observed in the data (e.g., test scores), one can infer features of the unobservable factors (e.g., their distributions).

While many methods exist to deal with endogenous treatment assignment in the static binary framework when there is selection on unobservables, such as difference-in-differences, regression discontinuity or instrumental variable approaches, these methods are often diffi-

[1] See also Bonhomme and Robin (2010) and Cunha, Heckman, and Schennach (2010) for recent developments.

cult to extend to a multiple heterogeneous treatment setting and considerably less research has focused on the case of multiple endogenous treatments.[2] Nekipelov (2008) provides a useful generalization of the instrumental variable approach to a multiple treatment setting. However, this relies on a monotonicity assumption that would be difficult to extend to our framework.

We build on the small literature on the analysis of treatment effects in dynamic models. Our approach to modelling time-varying treatments is close to that in Heckman and Navarro (2007). However, we focus on how factor analytic methods can aid in identification and interpretation of time-varying treatment effects. We generalize the factor structure results used in other settings (Carneiro, Hansen, and Heckman, 2003, Bonhomme and Robin, 2010). Our generalization is appealing not only because it is likely to be useful in other settings, but also because it is less data hungry, a common criticism of factor models. Furthermore, we link the assumptions used in factor structure models to better known fixed effects and regression discontinuity approaches.[3]

Our analysis is similar in spirit to that of Ham and LaLonde (1996) who point out the potential pitfalls of applying standard static methods to models where time since treatment matters, a problem we also study. However, as in Abbring and Van den Berg (2003), we also allow for an endogenously selected time of treatment to affect outcomes.[4] Unlike Ham and LaLonde (1996) and Abbring and Van den Berg (2003), our model does not rely on the proportional hazards assumption. Hence, we can allow for the unobservable to be multidimensional and for different unobservables to enter the model as time elapses. Furthermore, we permit the effect of the unobservables to change over time and with treatment status. Thus, our model supports more general forms of treatment heterogeneity than in either Ham and LaLonde (1996) (where treatment effects are homogeneous), or 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).

While there is a literature that allows for time at treatment to determine treatment effects (Gill and Robins, 2001; Murphy, 2003; Lechner, 2004), it is based on sequential conditional independence assumptions and so rules out selection on unobservables. Furthermore, Cellini, Ferreira, and Rothstein (2010) provide a useful generalization of the regression discontinuity approach to deal with dynamic selection, i.e., that individuals who are just below the threshold for treatment one period may be more likely to select into treatment the next. However, a key assumption for their model is that treatment effects do not vary by individual types.

[2] See Frölich (2004) for a discussion and Cataneo (2009) for recent developments.

[3] In Online Appendix C we provide further comparisons with other commonly employed methods.

[4] Abbring and Van den Berg (2003) show the nonparametric identifiability of a model similar to Ham and LaLonde (1996) in the context of a continuous time mixed proportional hazards model of duration.

In contrast, we consider the case where heterogeneity and selection based on unobservables are an important part of the problem. This type of heterogeneity is important in many cases. We find that these generalizations (i.e., the timing of treatment, heterogeneity in effects by unobservable student ability and selection on unobservables) play important roles in our application.[5]

We apply our method to study the effect of retention on achievement using data from the Early Childhood Longitudinal Study of Kindergartners (ECLS-K). We provide new insight into how treatment effects vary by students of different abilities as well as by the timing of and time elapsed since retention. We find evidence of dynamic selection and heterogeneous treatment effects by unobservable student abilities. For instance, while the average treatment effect of kindergarten retention is positive in the long run, the treatment effect on students who are actually retained in kindergarten is negative. We further find that the negative effect of retention on treated students generally diminishes as time since retention passes. In comparison, since the simpler fixed effect approach only provides at best estimates of the average treatment effect, it would lead to erroneous policy conclusions as the average student is not the typical student retained.

We also consider the effect of a change in retention policies to make it harder to retain students at the margin. Given considerable heterogeneity in treatment effects by ability, the effect for the marginal students is an important policy parameter, since the marginal student is very different from both the average student and the average student who is retained.

The paper proceeds as follows. In Section 2, we describe the basic framework and define treatment effects for the dynamic case. In Section 3, we specialize the framework to our proposed factor structure. We show that the model is semiparametrically identified. We discuss our application in Section 4.

2 The Framework

Consider the problem of evaluating the efficacy of a potentially time-varying treatment.[6] Let $t = 1, 2, ..., \bar{t}$ index calendar time and $i = 1, ..., I$ index the individual. Since we allow for the

[5] Furthermore, our focus is not on the design of optimal treatment regimes as in Murphy (2003) but rather on the identification of treatment effects. This distinction is important since the design of an optimal treatment regime, depending on the definition of optimality, may require the identification of different aspects than those we focus on. That is, while our results can be used to analyze and improve policy, they may not be enough (or may be more than required) to design a policy that satisfies particular requirements (i.e., to be "optimal" in some sense).

[6] Since in our empirical application we analyze the time-varying effects of grade retention, we continually use it as an example in the text to help fix ideas. There are many other examples in the literature that one could fit into this framework: how soon after pregnancy to stop smoking, when to participate in a training program for the unemployed, when to start (or stop) taking a drug, when to install a machine, etc.

treatment to be taken at different times (e.g., for children to be retained at different grades), we define a random variable R_i (whose realization we denote by r) that indicates the time at which treatment is received (e.g., the grade in which a student is retained). We assume that treatment is taken at most once.[7] We let $R_i = \{\underline{R}, \underline{R}+1, ..., \bar{R}-1, \bar{R}, \infty\}$, where $\underline{R} \geq 1$ and $\bar{R} \leq \bar{t}$ allows for the possibility that treatment can be taken only on a subset of the observed time periods. We adopt the convention of letting $R_i = \infty$ for the "never" treated state.[8]

The (possibly vector-valued) outcome of interest at time t for an individual i who receives treatment at time $R_i = r$ is denoted by $Y_i(t, r)$.[9] For notational simplicity, we keep all conditioning on covariates implicit. Finally, we define a random variable $D_i(r)$ that takes value 1 if an individual receives treatment at time r and 0 otherwise. For individual i the observed outcome in period t will be given by

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

As opposed to the standard binary treatment case, we now have many possible potential outcomes. That is, while the standard case only has the treated and untreated potential states, we have the untreated, the treated at time $\underline{R}$, the treated at time $\underline{R}+1$, etc. Because of the sequential nature of the problem, by letting $Y_i(t, r)$ depend on treatment time r, we allow for the possibility that the effect of treatment depends not only on receipt but on the time at which treatment is received. For example, there is no single effect of retention, but rather an effect of retention in kindergarten, in first grade, etc. Furthermore, there is no single effect of retention in kindergarten (for example), as the effects depend on the time elapsed since retention. This setting can also be interpreted as depending on the time since treatment $(t - r)$, making it straightforward to analyze the outcomes as durations, counts, etc.

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

A-1 $Y_i(t, r) = Y_i(t, \infty) = Y_i(t)$ *for* $r \geq t$ *and* $r \neq \infty$.

That is, we rule out that potential outcomes differ because *in the future* treatment times will be different. In our application this means, for example, that after conditioning on all

[7] Extending the framework to allow for the possibility of treatment being taken more than once can be done at the cost of introducing a lot more notation, by letting R_i be a random vector characterizing the times at which an individual receives treatment.

[8] Depending on the situation this case may be more accurately described as the "not treated yet" or "not treated in the sample period."

[9] These could be a vector of continuous test scores given a retention status (as in our application), a vector of discrete random variables (measuring attendance for example), strings of discrete random variables (as in a duration model, time until graduation for example) or combinations of these.

prior information, the fact that a student will be retained in second grade does not directly affect her performance in first grade. While Abbring and Van den Berg refer to this as the *no anticipations* assumption, importantly this should not be confused with the assumption that individuals are not forward looking. Assumption **A-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).[10]

We further write the outcomes as

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

where, because of **A-1**, we impose $\Phi(t,r) = 0$ and $\epsilon_i(t,r) = \epsilon_i(t)$ if $r \geq t$ *and* $r \neq \infty$.[11]

We assume that selection into treatment and treatment time are determined by a single spell duration model that follows a sequential threshold crossing structure as in Heckman and Navarro (2007). If we define the treatment time specific index $V_i(r) = \lambda(r) + U_i(r)$, then treatment time is selected according to

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

where $\mathbf{1}(a)$ is an indicator function that takes value 1 if a is true and 0 otherwise. The selection process is dynamic in the sense that today's choice 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. In many situations it is not clear how to fully specify the selection process. Our application provides a good example of this, since the decision to have a student repeat a grade is the result of some complex process involving many actors. Our analysis provides an alternative to the fully-specified structural model by extending the standard selection model to account for dynamics. Cunha, Heckman, and Navarro (2007)

[10] What it rules out is 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 **A-1** to hold. We can accommodate cases in which **A-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.

[11] We treat the outcome as continuous for convenience. We can easily work with discrete and mixed discrete/continuous outcomes by defining them as random variables arising from other latent variables crossing thresholds. For example, if the outcome were binary, we can define a latent variable $Y_i^*(t,r) = \Phi(t,r) + \epsilon_i(t,r)$ so that the measured outcome $Y_i(t,r)$ would be $Y_i(t,r) = \mathbf{1}(Y_i^*(t,r) > 0)$ where the function $\mathbf{1}(a)$ takes value 1 if a is true and 0 if it is not. Furthermore, additive separability in outcomes is not strictly required, it can be relaxed using the analysis in Matzkin (2003).

provide conditions under which structural dynamic discrete choice models can be represented by a reduced form approximation as above. Furthermore, since extending it to the case in which treatment is not an absorbing state (i.e., treatment can be received more than once) is straightforward it can be applied in more complex situations. [12]

The observed outcome in period t is then given by

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

If there is no selection based on unobservables, then the problem is easier and we can recover an unbiased estimate of the effect of treatment on outcomes. In general this is not the case, and some of the same unobservables that determine the outcome determine the selection process. For instance, higher ability students may be less likely to be retained and more likely to have higher test scores. Our goal in general is to allow $\left(\epsilon_i\left(t,r\right),\epsilon_i\left(t',r''\right),U_i\left(r'''\right),U_i\left(r''''\right)\right)$ all to be correlated.

2.1 Defining Treatment Effects

Before turning to the identification problem, we first consider the problem of defining what constitutes "the" effect of treatment at the individual level. We can define at least two different candidates for the individual effect of treatment. The first parameter

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

measures the effect at period t of receiving treatment at time r versus receiving treatment at time r'. 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. An example of this first parameter would be the difference in test scores at age 11 for a student if he repeats first grade versus if he repeats third grade.

The second individual parameter of interest

$$\begin{aligned}\Delta_i^2\left(\tau,r,r'\right) &= Y_i\left(r+\tau,r\right)-Y_i\left(r'+\tau,r'\right)\\ &= \Phi\left(r+\tau,r\right)-\Phi\left(r'+\tau,r'\right)+\epsilon_i\left(r+\tau,r\right)-\epsilon_i\left(r'+\tau,r'\right),\ \text{for } \tau>0\end{aligned}$$

[12] In this case, we would generalize the threshold crossing model into a multiple spell model, where the whole sequence of prior treatments/no treatments potentially affects the decision each period. R_i would be a vector containing the treatment history up to t, and an individual would choose treatment every time the index becomes positive (not only the first time).

measures the difference in the effect of receiving treatment τ periods after treatment time for two different treatment times r and r'. An example of this parameter would be the difference in wages one year after taking a training program if the individual takes the training 3 months after the unemployment spell starts versus 6 months after the spell begins.

Regardless of how one defines the effect of treatment, we can consider what happens as time since treatment elapses. The effect is potentially individual specific even conditional on covariates. Relative to the static binary case, in the time-varying setting there are many more possible population average parameters, both of the average treatment effect and treatment on the treated type. For example, we can define the average effect of receiving treatment at time $R_i = r$ versus not receiving treatment

$$ATE\left(t,r\right) = E\left(Y\left(t,r\right) - Y\left(t,\infty\right)\right) = \Phi\left(t,r\right) - \Phi\left(t,\infty\right);$$

the average effect of treatment at time t for people who receive treatment at time $R_i = r$

$$TT\left(t,r\right) = E\left(Y\left(t,r\right) - Y\left(t,\infty\right)|R_i = r\right)$$

and so on. In our example, this could be the average effect on third grade test scores of being retained in kindergarten versus not being retained, for those children who were retained in kindergarten. Because of the multiplicity of treatments available, we can define many more mean treatment parameters like the average effect of receiving treatment at $R_i = r$ versus receiving treatment at $R_i = r'$

$$ATE\left(t,r,r'\right) = E\left(Y\left(t,r\right) - Y\left(t,r'\right)\right)$$

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

$$TT\left(t,r,r',r''\right) = E\left(Y\left(t,r\right) - Y\left(t,r'\right)|R_i = r''\right),$$

etc. For instance, we may want to know the return to retaining students in kindergarten who were actually retained in first grade.

In general, depending on whether we assume the mean component $\Phi\left(t,r\right)$ and/or the unobserved component of the outcome $\epsilon_i\left(t,r\right)$ depend on r or not, the effect of treatment is time-varying. In the same manner, depending on whether $\epsilon_i\left(t,r\right)$ varies across individuals, the effect is heterogeneous in the population. Under certain assumptions that limit the heterogeneity of treatment effects some of these parameters may equal one another. We focus on the more general case, where the treatment effect is allowed to vary over time and by unobserved individual characteristics. Both of these types of heterogeneity prove important in our application.

3 Identification

The primary challenge to identifying treatment effects in the static framework is that individuals differ in unobservable ways that help determine both selection into treatment and the effect of treatment. 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 dynamic setting, with the added challenge that selection is dynamic and that treatment effects vary both by the unobservable type of the individual and over time.

In this section, we develop a methodology based on a factor-analytic approach for dealing with dynamic selection and heterogeneous, time-varying treatment effects. We focus on the case where returns are heterogeneous both because this case is arguably empirically more relevant and because applying standard instrumental variables methods under homogeneity of treatment effects is a straightforward GMM problem.[13] We then describe conditions such that the model is semiparametrically identified.

Our approach can be understood as a hybrid between the control function and a generalized version of the fixed effect approach. As with all control function based methods, identification is more transparent and easier to achieve when instruments are available, but they are not strictly required. In contrast to the standard fixed effect approach, we can allow for the individual effects to be multidimensional, time-varying and treatment-specific (e.g., the effect of ability can differ in the retained relative to the non-retained states).[14]

3.1 Factor Structure

For illustration, consider a simple 3 period example where treatment can be taken in either of the first 2 periods ($R = 1, 2$), e.g., students can be retained in kindergarten or first grade. The policy is evaluated according to its effect on some ex-post outcome measured at period t: $Y_i\left(t, r\right)$, e.g., third grade test scores. For example, the potential outcomes in period 3 can

[13] Notice that, because of the dynamic nature of the model, even if we only have one instrument Z, but it is time varying it can potentially be used as an instrument for all $D_i\left(r\right)$ since the choices are made sequentially over time.

[14] In Online Appendix C we briefly discuss some of the advantages and shortcomings of applying commonly employed approaches in the static treatment literature in our dynamic setting.

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

$$\begin{aligned} Y_i(3) &= \Phi(3,\infty) + D_i(1)[\Phi(3,1) - \Phi(3,\infty)] + D_i(2)[\Phi(3,2) - \Phi(3,\infty)] \\ &\quad + \epsilon_i(3,\infty) + D_i(1)[\epsilon_i(3,1) - \epsilon_i(3,\infty)] + D_i(2)[\epsilon_i(3,2) - \epsilon_i(3,\infty)]. \end{aligned} \tag{3}$$

The (observed) outcome equation in period 3 is a regression model with dummy indicators for the time at which an individual receives treatment. Notice that this is not a standard binary treatment model both because we now have more than one treatment indicator and because the effect of treatment is potentially heterogeneous. In the language of Heckman, Urzua, and Vytlacil (2006), we have a situation in which essential heterogeneity is present if the decision of when to receive treatment is correlated with the unobservable (to the econometrician) gains of choosing each treatment. That is, in our case $D_i(r)$ and/or $D_i(r')$ are likely to be correlated with $\epsilon_i(3,r) - \epsilon_i(3,r')$ for $r \neq r'$. In the retention example, essential heterogeneity exists if the students who are retained are more likely to experience higher (lower) gains from retention.

One way to account for essential heterogeneity is to identify and estimate the joint distribution of all the unobservables (U_i, ϵ_i) . This would permits us to 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 assume:

A-2 *(Factor structure)* $\epsilon_i(t,r) = \theta_i\alpha(t,r) + \varepsilon_i(t)$ and $U_i(r) = \theta_i\rho(r) + \upsilon_i(r)$ where θ_i is a vector of mutually independent "factors" and we assume that $\varepsilon_i(t) \perp\!\!\!\perp \varepsilon_i(t')$ for all $t \neq t'$, $\upsilon_i(r) \perp\!\!\!\perp \upsilon_i(r')$ for all $r \neq r'$ and $\upsilon_i(r) \perp\!\!\!\perp \varepsilon_i(t)$ for all r and t where $\perp\!\!\!\perp$ denotes statistical independence.[15]

We impose **A-2** for convenience, even though it is stronger than required.[16] The factor structure assumption is a convenient dimension reduction technique: it reduces 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 $\varepsilon_i(t), \upsilon_i(r)$ $\forall t, r$.

The factor structure also has an appealing interpretation, since we can now talk about

[15] If **A-1** holds, $\alpha(t,r) = \alpha(t,\infty) = \alpha(t)$ for $r \geq t$.

[16] 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.

a low dimensional set of common "causes."[17] The same set of unobservables (the vector θ_i) that determines the effect of treatment also determines selection into treatment. In our grade retention example, if θ_i is a vector of unobserved "abilities," essential heterogeneity arises because unobserved ability affects both the gain in test scores across two years and the probability of being retained. We can then consider questions such as whether less able students in our model are more likely to be retained earlier or later and test the implications for the effect of treatment on these students.

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

$$V_i\left(r\right) = \lambda\left(r\right) + \theta_i \rho\left(r\right) + \upsilon_i\left(r\right).$$

The observed outcomes are

$$Y_i\left(1\right) = \Phi\left(1\right) + \varepsilon_i\left(1\right) + \theta_i \alpha\left(1\right),$$

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

and

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

In this case, essential heterogeneity is present when $\alpha\left(3,r\right) \neq \alpha\left(3,\infty\right)$ or $\alpha\left(2,r\right) \neq \alpha\left(2,\infty\right)$, since now the unobserved gains in the test score

$$\epsilon_i\left(t,r\right) - \epsilon_i\left(t,\infty\right) = \theta_i\left[\alpha\left(t,r\right) - \alpha\left(t,\infty\right)\right]$$

are correlated with the choice indicator because the same θ_i determines both.

If we could recover (or condition on) the unobserved θ_i, then $D_i(1)$ and $D_i(2)$ are no longer endogenous and we can obtain consistent estimates 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. There are many normalizations under which the distribution of θ_i can be recovered (see Cunha, Heckman, and Schennach, 2010 and Bonhomme and Robin,

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

2010 for examples).

To understand how the factor model we propose attempts to generalize the fixed effect model, take differences between the period 2 and period 1 outcomes to difference out the individual effect θ_i, so that

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

For the differencing strategy to work we need to impose two restrictions. First, we would need to rule out essential heterogeneity, i.e., $\alpha(2,1) = \alpha(2,\infty) = \alpha(2)$. Second, we would additionally have to assume that the marginal effect of θ_i does not change over time so $\alpha(2) = \alpha(1) = \alpha$. First differencing eliminates θ_i only when these two restrictions hold. As more periods pass, more assumptions are required for the fixed effect model to work. For instance, to identify the effect on period 3 outcomes, we would need to impose the additional assumption that $\alpha(3,2) = \alpha(3,1) = \alpha(3,\infty) = \alpha(3)$.

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 $\Phi(3,1) - \Phi(3,\infty) - 2(\Phi(2,1) - \Phi(2,\infty))$ so we cannot separate the effect of being treated in period 1 on outcomes in periods 2 and 3. Note that under the assumptions that make the differencing strategy possible, the average treatment effect is the same as the treatment on the treated. In many cases, including our application, this is not a reasonable assumption. Hence, using an identification strategy that allows for essential heterogeneity is important.

The main goal of the factor structure, as we propose it, is to allow for the possibility of essential heterogeneity, multidimensional abilities and the marginal effects of abilities to vary by treatment status. To illustrate how the factor structure works, consider a simple 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, i.e., the standard case in which one assumes that unobserved ability is uni-dimensional. Suppose the outcome in period 1 is free of selection,[18] so

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

[18] Alternatively if we have access to an exclusion restriction (i.e., an instrumental variable) we can control for selection nonparametrically as in Heckman (1990) and Heckman and Smith (1998) and work with selection corrected outcomes.

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). From it, normalizing $\rho_1(1) = 1$,[19] we can form

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

With $\alpha_1(1)$ in hand it follows from a Theorem of Kotlarski (1967)[20] that the distribution of $\theta_{i,1}$ (and of $\varepsilon_i(1)$ and $\upsilon_i(1)$) is nonparametrically identified. 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 $E\left(\epsilon_i^k(1)\,U_i(1)\right) = \alpha_1^k(1)\,E\left(\theta_{i,1}^{k+1}\right)$ for $k > 0$. Formally, one wants to characterize a distribution using its characteristic function and not moments, and this is precisely what the Kotlarski argument does.

Next consider the (selection corrected) second period outcomes

$$Y_i(2,r) = \Phi(2,r) + \theta_{i,1}\alpha_1(2,r) + \theta_{i,2}\alpha_2(2,r) + \varepsilon_i(2) \text{ for } r \in \{1,\infty\}$$

and selection equation

$$V_i(2) = \lambda(2) + \theta_{i,1}\rho_1(2) + \theta_{i,2}\rho_2(2) + \upsilon_i(2)\,,$$

where we now allow for a new element of θ_i $(\theta_{i,2})$ to enter the model. $\theta_{i,2}$ can be interpreted as a correlated shock, i.e., an unobserved shock that affects outcomes and selection equations from period 2 onward, with the potential that its effect may change as time elapses. Alternatively, one can think of it as an ad-hoc way of letting unobserved ability evolve over time. By taking cross moments over time (i.e., $Y_i(1)$ with the selection corrected $Y_i(2,r)$), we can identify the elements associated with $\theta_{i,1}$ in period 2 equations. Then, by taking cross moments within period 2 equations, we can identify the elements associated with the correlated shock $(\theta_{i,2})$, as well as the nonparametric distributions of the unobservables.

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

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

$$\begin{aligned} Z_1 &= X_1 - X_2 \\ Z_2 &= X_1 - X_3; \end{aligned}$$

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.

We extend this analysis to the case in which unobserved ability (θ_i) is multidimensional beyond the correlated shocks (i.e., gaining a new element of θ_i each period). Associated with ability is a set of tests or markers that measure these components of ability imperfectly. In our empirical example, these correspond to the initial tests given to students in kindergarten before any grade repetition takes place. The existence of selection-free initial test scores is not crucial (provided we can correct for selection), but we keep it because a) it is common to many situations and b) it simplifies the exposition of the identification argument.[21]

In our empirical application we consider a normalization of θ_i that is particularly relevant to retention decisions we propose that true ability at the initial period consists of three independent components (A_i, B_i, C_i). In particular, assume we have 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 j^{th} demeaned cognitive test as

$$\zeta_{i,j} = A_i \alpha_{\zeta,j} + C_i \pi_{\zeta,j} + \varepsilon_{i,\zeta,j}, \tag{4}$$

and the j^{th} demeaned behavioral test as

$$\beta_{i,j} = A_i \alpha_{\beta,j} + B_i \phi_{\beta,j} + \varepsilon_{i,\beta,j}. \tag{5}$$

Under this interpretation, tests are noisy measures of the components of ability. Depending on the nature of the measure, some (like math and reading test scores) are markers of cognitive ability C_i and general ability A_i and some (like measures of class disruptive behaviors or habits) are noisy measures of the behavioral ability B_i and general ability A_i. 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. Other normalizations are possible, but the present normalization may also be applicable to other settings with multidimensional unobservables.

Semiparametric identification follows similarly to the one factor model. Now we take moments across cognitive and behavioral equations to recover the α parameters and the nonparametric distribution of A. We then take cross moments within cognitive tests and within behavioral tests to recover the π and ϕ parameters, as well as the nonparametric

[21] There is nothing special about ability and tests. In a different setting, we could refer to abilities as general and specific unobservables, and to test scores as measurements. For ease of exposition, however, we continue referring to these unobserved factors as general, behavioral and cognitive abilities and to the measurements associated with them as test scores.

distributions of B, C and the $\varepsilon' s$.

Formally, without loss of generality, we impose the following normalizations $\alpha_{\zeta,1} = 1$, $\pi_{\zeta,1} = 1$ and $\phi_{\beta,1} = 1$.[22] We first take cross moments between cognitive and behavioral measures

$$\begin{array}{c} E\left(\left(\zeta_j\right)^n \beta_k\right) = \alpha_{\zeta,j}^n \alpha_{\beta,k} E\left(A^{1+n}\right) \\ E\left(\zeta_j \left(\beta_k\right)^m\right) = \alpha_{\zeta,j} \alpha_{\beta,k}^m E\left(A^{1+m}\right) \end{array}. \tag{6}$$

and form

$$\frac{E\left(\zeta_j \left(\beta_k\right)^n\right)}{E\left(\zeta_1 \left(\beta_k\right)^n\right)} = \frac{\alpha_{\zeta,j} \alpha_{\beta,k}^n E\left(A^{1+n}\right)}{\alpha_{\beta,k}^n E\left(A^{1+n}\right)} = \alpha_{\zeta,j}$$

to recover all of the general ability loadings on cognitive tests, $\alpha_{\zeta,j}$, for $j = 2, \ldots, N_c$. We can then, for example, form

$$\frac{E\left(\zeta_1 \left(\beta_k\right)^2\right)}{E\left(\left(\zeta_1\right)^2 \beta_k\right)} = \frac{\alpha_{\beta,k}^2 E\left(A^3\right)}{\alpha_{\beta,k} E\left(A^3\right)} = \alpha_{\beta,k}$$

and recover the general ability loadings on behavioral tests.

To show that the distribution of A is identified, without loss of generality, take any two tests, for example a cognitive and a behavioral one, and form

$$\frac{\zeta_{i,j}}{\alpha_{\zeta,j}} = \left[C_i \frac{\pi_{\zeta,j}}{\alpha_{\zeta,j}} + \frac{\varepsilon_{i,\zeta,j}}{\alpha_{\zeta,j}}\right] + A_i,$$

$$\frac{\beta_{i,k}}{\alpha_{\beta,k}} = \left[B_i \frac{\phi_{\beta,k}}{\alpha_{\beta,k}} + \frac{\varepsilon_{i,\beta,k}}{\alpha_{\beta,k}}\right] + A_i.$$

Then, using Kotlarski (1967), the distribution of A (and of $\left[C \frac{\pi_{\zeta,j}}{\alpha_{\zeta,j}} + \frac{\varepsilon_{\zeta,j}}{\alpha_{\zeta,j}}\right]$ and $\left[B \frac{\phi_{\beta,k}}{\alpha_{\beta,k}} + \frac{\varepsilon_{\beta,k}}{\alpha_{\beta,k}}\right]$) is nonparametrically identified.

With all of the parameters associated with general ability A as well as its distribution identified, we can then take the system of cognitive tests and form

$$E\left(\zeta_j \left(\zeta_k\right)^n\right) - \alpha_{\zeta,j} \alpha_{\zeta,k}^n E\left(A^{1+n}\right) = \pi_{\zeta,j} \pi_{\zeta,k}^n E\left(C^{1+n}\right),$$

for any $j \neq k$ with $j, k = 1, ..., N_c$. By, for example, forming

$$\frac{E\left(\zeta_1 \left(\zeta_k\right)^2\right) - \alpha_{\zeta,1} \alpha_{\zeta,k}^2 E\left(A^3\right)}{E\left(\left(\zeta_1\right)^2 \zeta_k\right) - \alpha_{\zeta,1}^2 \alpha_{\zeta,k} E\left(A^3\right)} = \frac{\pi_{\zeta,k}^2 E\left(C^3\right)}{\pi_{\zeta,k} E\left(C^3\right)} = \pi_{\zeta,k},$$

we can recover $\pi_{\zeta,k}$ for $k = 2, ..., N_c$. By iteratively applying the Kotlarski argument, we

[22] Given that A, B, and C are all latent, these normalizations imply no restriction since $A\alpha_{\zeta,j} = A\kappa \frac{\alpha_{\zeta,j}}{\kappa}$ for any constant κ.

can nonparametrically recover the distributions of C and $\varepsilon_{\zeta,j}$ for $j = 1, ..., N_c$. Finally, by applying the same argument to the system of behavioral tests, we can recover $\phi_{\beta,j}$ and the nonparametric distributions of B and $\varepsilon_{\beta,j}$ for $j = 1, ..., N_b$.

Once we have recovered the distribution of (A_i, B_i, C_i), we can proceed to the next period. Now some children will be treated (i.e., will repeat kindergarten), and so the test scores in period 2 will be contaminated with selection. By using the selection equation, we can correct period 2 test scores using semiparametric selection correction methods like the control function approach.[23] We can then repeat the arguments above and recover the loadings and the distribution of the $\varepsilon's$. 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. The normalization that B_i only enters β equations and C_i only enters ζ equations need only apply to the first period. By proceeding iteratively, we can recover all of the outcomes of interest.

Here we assume that the only determinants of selection are the A, B, C components of ability. Since we can identify those elements in period 1, we can add new elements to θ over time to allow for new persistent unobserved (to the econometrician) shocks every period, as in the example where ability is single-dimensional.

Formally, consider a modified version of the model of equations (4) and (5) in a multi-period setting. In period 1 the model is given by:

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

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

Identification of these period 1 equations follows exactly as before. Moving forward in time we have that the demeaned selection corrected period t cognitive tests for retention status r are written as

$$\zeta_{i,j,r,t} = A_i \alpha_{\zeta,j,r,t} + B_i \phi_{\zeta,j,r,t} + C_i \pi_{\zeta,j,r,t} + \sum_{\tau=2}^{t} \eta_i^{(\tau)} \delta_{\zeta,j,r,t}^{(\tau)} + \varepsilon_{i,\zeta,j,t}. \qquad (7)$$

First, notice that we now allow for behavioral ability to determine cognitive tests after period 1. Second, we also add a new unobservable $\eta_i^{(\tau)}$ every period. Since this new unobservable is individual specific and affects all outcomes (and retention decisions) from period τ on, it

[23] Notice that the selection equation in period 1 only depends on (A_i, B_i, C_i) and so, strictly speaking, an exclusion restriction is not required for nonparametric identification as in Heckman (1990) and Heckman and Smith (1998). See Heckman and Robb (1985) and Navarro (2008) for use of control functions to control for selection.

can be interpreted as a permanent shock that first affects outcomes in period τ (hence the superscript). While the shock itself is permanent, we allow for its effects to change both over time and across retention statuses for all equations in the model.

Now consider identification of equation (7) in period 2 for an arbitrary retention status r. We can form cross second moments between period 2 and period 1 cognitive tests:

$$\begin{aligned} E\left(\zeta_{j,r,2},\zeta_{j,1}\right) &= \alpha_{\zeta,j,r,2}\left[\alpha_{\zeta,j,1}E\left(A^2\right)\right]+\pi_{\zeta,j,r,2}\left[\pi_{\zeta,j,1}E\left(C^2\right)\right] \\ E\left(\zeta_{j,r,2},\zeta_{k,1}\right) &= \alpha_{\zeta,j,r,2}\left[\alpha_{\zeta,k,1}E\left(A^2\right)\right]+\pi_{\zeta,j,r,2}\left[\pi_{\zeta,k,1}E\left(C^2\right)\right]. \end{aligned}$$

The terms in square brackets are all known from our period 1 analysis. Provided a standard rank condition holds, this system can be solved for both $\alpha_{\zeta,j,r,2}$ and $\pi_{\zeta,j,r,2}$ for $j=1,...,N_c$. Then, by taking cross second moments with period 1 behavioral tests we can form:

$$\frac{E\left(\zeta_{j,r,2},\beta_{k,1}\right)-\alpha_{\zeta,j,r,2}\left[\alpha_{\beta,k,1}E\left(A^2\right)\right]}{\phi_{\beta,k,1}E\left(B^2\right)}=\phi_{\zeta,j,r,2}$$

and recover the behavioral ability loadings $\phi_{\zeta,j,r,2}$ for $j=1,...,N_c$.

In order to identify the terms related to the new unobservable (i.e., the period 2 permanent shock $\eta^{(2)}$ and its loadings $\delta^{(2)}_{\zeta,j,r,2}$), a normalization on the scale of the unobservable is required. We impose that $\delta^{(2)}_{\zeta,1,\infty,2}=1$. We form cross moments between period 2 equations for the $r=\infty$ retention status and get

$$\frac{\left[\begin{array}{c} E\left(\zeta_{j,\infty,2},\zeta_{k,\infty,2}\right)-\alpha_{\zeta,j,\infty,2}\alpha_{\zeta,k,\infty,2}E\left(A^2\right) \\ -\phi_{\zeta,j,\infty,2}\phi_{\zeta,k,\infty,2}E\left(B^2\right)-\pi_{\zeta,j,\infty,2}\pi_{\zeta,k,\infty,2}E\left(C^2\right) \end{array}\right]}{\left[\begin{array}{c} E\left(\zeta_{1,\infty,2},\zeta_{k,\infty,2}\right)-\alpha_{\zeta,1,\infty,2}\alpha_{\zeta,k,\infty,2}E\left(A^2\right) \\ -\phi_{\zeta,1,\infty,2}\phi_{\zeta,k,\infty,2}E\left(B^2\right)-\pi_{\zeta,1,\infty,2}\pi_{\zeta,k,\infty,2}E\left(C^2\right) \end{array}\right]}=\delta^{(2)}_{\zeta,j,\infty,2}$$

to identify the loadings on the permanent shock for all cognitive scores $j=1,...,N_c$ and retention status $r=\infty$.[24] We can then apply Kotlarski to any pair of equations j,k for $r=\infty$ and identify the nonparametric distributions of $\eta^{(2)}$ and $\varepsilon_{\zeta,j,2},\varepsilon_{\zeta,k,2}$. To identify the loadings for retention statuses $r\neq\infty$, we can form

$$\frac{\left[\begin{array}{c} E\left(\zeta_{j,r,2},\zeta^2_{k,r,2}\right)-\alpha_{\zeta,j,r,2}\alpha^2_{\zeta,k,r,2}E\left(A^3\right) \\ -\phi_{\zeta,j,r,2}\phi^2_{\zeta,k,r,2}E\left(B^3\right)-\pi_{\zeta,j,r,2}\pi^2_{\zeta,k,r,2}E\left(C^3\right) \end{array}\right]}{\left[\begin{array}{c} E\left(\zeta_{j,r,2},\zeta_{k,r,2}\right)-\alpha_{\zeta,j,r,2}\alpha_{\zeta,k,r,2}E\left(A^2\right) \\ -\phi_{\zeta,j,r,2}\phi_{\zeta,k,r,2}E\left(B^2\right)-\pi_{\zeta,j,r,2}\pi_{\zeta,k,r,2}E\left(C^2\right) \end{array}\right]}\frac{E\left(\left(\eta^{(2)}\right)^2\right)}{E\left(\left(\eta^{(2)}\right)^3\right)}=\delta^{(2)}_{\zeta,k,r,2}.$$

[24] Notice that we cannot form cross moments for equations with different retention indices r, since we can only observe a student in the retention status he actually receives.

Applying the same arguments recursively, it is clear that we can add a new permanent shock every period and still identify all of the loadings and nonparametric distributions of the unobservables. The factor structure has other advantages. For example, we can correct for potential biases due to selective sample attrition (e.g., children moving to a different school if they know they will be retained in their current school) by adding an equation for missing data (say a binary model for attrition) that depends on the same common vector θ_i .

4 The Effect of Retention on Test Scores

Most research on the effects of grade retention treats it as a single treatment (being retained versus not being retained) by focusing on the outcomes at a single grade. These studies generally find that retention at best has no effect and at worse has considerable negative effects.[25] In some recent studies, Jacob and Lefgren (2004) and Nagaoka and Roderick (2005) use a regression discontinuity design to study test-based promotion in Chicago public schools that applied to third and sixth graders. Both studies find that retention leads to small short term gains on test scores for third graders, but no effects on sixth graders. This disparity in the estimated effect of retention across third and sixth grade, however, could also follow if the marginal student differs in unobservable ways across grades (i.e., dynamic selection).

Another study by Jacob and Lefgren (2009) considers long run effects using a similar design and finds no effects on high school completion for students who were retained in sixth grade under the Chicago policy. As discussed by Cellini, Ferreira, and Rothstein 2010 in their implementation of a dynamic RD design, a difficulty with evaluating long run effects in this framework is that students who are just above the margin for passing in sixth grade (i.e., the control group) may also be more likely to be retained later, contaminating the control group. Thus, while these studies provide important insight into the effects of grade retention, they can only tell us about the effect of retention at the certain grades where the policy applies, for students on the margin of being retained under that policy, and for a particular type of long run effect.[26] We bring new insight to this literature by estimating how the effect of grade retention varies across different grades, as time since retention passes and by different types of students.

We study the effects of grade retention using the ECLS-K, a nationally representative survey of kindergartners in 1998/99. It follows the students as they progress through school,

[25] See Holmes (1989) and Jimerson (2001) for comprehensive meta-analyses.

[26] Furthermore, as the authors discuss, these studies focus on a setting where retention is linked to high stakes testing, and is thus associated with a set of incentives for teachers and students that may not apply to the effect of retention in other settings.

with follow-up surveys in the 1999/2000, 2001/02 and 2003/04 school years. A benefit of this data set is that we observe the whole 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 only once, did not skip grades, and were taking kindergarten for the first time in 1998/99.[27] 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).[28] That is, our dynamic treatment time indicator takes values $R_i = 1, 2, 3, \infty$, where $R_i = \infty$ means the child is never retained, $R_i = 1$ that he is retained in kindergarten, $R_i = 2$ that he is retained early and $R_i = 3$ that he is retained late.

Each year of the ECLS-K includes cognitive tests measuring students' science,[29] reading and math skills. 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. For behavioral measures we use teacher ratings on students' behavioral and social skills—the approach to learning, self-control and interpersonal skills components of the Social Rating Scale (SRS).

A logical difficulty in evaluating the effect of grade retention is that it is impossible to hold both the grade and age fixed when determining the gains in achievement for a retained student. 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 would address, for instance, whether a student learns more by the end of fifth grade than he would if he had not repeated fourth grade. This would attribute maturation (or age) effects to the estimated effect of grade retention. Alternatively, holding age fixed would measure whether a student learns more, say, by age 11 if he repeats fourth grade than he would have if he had been promoted to the fifth grade and exposed to new material. We focus on the effect of retention holding age fixed, which the test scores in the ECLS-K are better-suited for measuring.

The ECLS-K contains a very rich set of covariates. We use characteristics of the children, the family, the class and the school as controls in our model. Class and teacher characteristics are taken from teacher surveys.[30] School administrator surveys provide information about

[27] The number of students who 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.

[28] In principle we could separate early and late into the four grades at which retention takes place. This, however, can only be done for less than half of the sample.

[29] 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.

[30] For the 2003/04 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

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 2 to 4. We restrict the sample to 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. We do this so that we can include as much of the data as possible in estimating the different outcome equations. A potentially important concern with a panel study of this type is non-random sample attrition. Column 6 of Table 1 shows the mean 1998/99 characteristics for students who are still in the sample in 2003/04 (the last year of the survey that we use for estimation). The number of observations decreases substantially across these years, from 7832 in the base year to 2106 in the last year. Comparing summary statistics, we see suggestive evidence of non-random attrition. For instance, 77% of the participants are white in 2003/04 compared to only 65% in 1998/99. Students in 2003/04 sub-sample also have higher SES and come from schools/classrooms with lower percentages of minorities than the initial sample. Our estimator controls for non-random sample attrition, as discussed in Section 4.1.

Table 2 describes how characteristics vary by retention statuses.[31] A total of 630 students in our base year sample are retained either in kindergarten, early or late. Only 87 are retained late, whereas 255 and 288 are retained in kindergarten and early, respectively. Retained students have lower average test scores than non-retained students, though it is less pronounced for reading tests. Among retained students, early retainees stand out as having the lowest test scores. Males are more likely to be retained than females, particularly in kindergarten and early. Nonwhite students are more likely to be retained than whites. Furthermore, students who are retained early or late are more likely to be nonwhite than students who are retained in kindergarten. In particular, 36% of kindergarten retainees are black or Hispanic, compared to 48% for early retainees and 46% for late retainees. Comparisons are similar across other characteristics, with retained students generally facing lower school and teacher quality, particularly students retained early or late. These summary statistics suggest that students not only differ in observables across whether they are retained or not, but also by when they are retained, providing some motivating evidence that dynamic selection on unobservables may also be a concern.

The ECLS-K also includes information on the schools' retention policies for the 1998/99, 1999/00 and 2001/02 survey years. We use these variables as exclusions, under the assump-

classroom measures for each test in estimating the outcome equations.

[31] Note that the total number of students added across categories is smaller than that reported in Table 1 (7668 compared to 7832). This is because not all students have retention indicators. However, we can still use test observations for students with missing retention indicators in the initial period to calculate our factors.

tion that, conditional on the other covariates including observable school characteristics, they do not directly determine the child's test score but they do affect the probability that a child repeats a grade. These policies include whether the school has a policy that allows children 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 multiple times in a given grade. As shown in Table 2, retention policies vary considerably across schools and also to a less 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 1998/1999 school year 44% of schools in the non-retained sample permit retention without parental permission, compared to 61% and 58% for students who are retained early or late.

While the summary statistics provide some suggestive evidence of dynamic selection, we further investigate selection in the raw data and potential evidence of time-varying treatment effects in Table 3. 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 child is retained in the future. We also control for covariates related to the child, his family, school and class, as described in Table 1 above. Column 2 of Table 3 presents results for reading and math in Panels A and B respectively. Not surprisingly, children who will be retained have lower kindergarten test scores than those who will not be retained. Furthermore, we reject the hypothesis that the coefficients on being retained at different grades in the future are the same. The p-values for these joint tests are included at the bottom of the table. 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 and late retainees respectively. These results suggest not only the presence of selection but also *dynamic* selection on cognitive test scores in the sense that different types of students are being retained at different grades.

We then look for evidence of time-varying treatment effects by regressing test scores in the last sample period (2003/04 school year) on retention in different grades. As shown in column 3 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, i.e., different types of students being retained in different grades.

One way to begin to control for a static component of selection is to include various performance measures in kindergarten, prior to any retention decisions taking place. Columns

4 and 5 present results controlling for kindergarten cognitive test scores and then behavioral test 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 can reject the formal test of equality of the effects for different retention times, again providing evidence for potentially time-varying treatment effects.[32] After including all initial test controls, retention in kindergarten is estimated to lower achievement by 9%, early retention by 14% and late 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, under our interpretation of tests scores as noisy measures of true latent abilities, using the kindergarten measures as controls may actually worsen the bias in the estimated treatment effects.[33] Furthermore, this analysis does not capture heterogeneous effects of treatment by student type, which is a central contribution of our paper.

4.1 Estimating a Multidimensional Model of Ability and Retention

We follow the discussion of identification in Section 3.1 and 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.[34] 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.

Let $\zeta_{i,j,1}$ be our j^{th} cognitive measure for individual i in period 1 (kindergarten) and similarly for behavioral measures. Our kindergarten measures are modeled as

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

and

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

[32] The same pattern holds for the other cognitive tests and behavioral measures.

[33] See Heckman and Navarro (2004).

[34] In all cases our cognitive test scores are measured as the log of the IRT scores, while our behavioral measures are defined to be the standardized SRS scores.

Our model for test scores in the following years is given by

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

We restrict the observable covariates (except for the constant) to have the same marginal effect across time for a given subject. We also restrict the effect of the permanent shock (η_i^{τ}) to be the same regardless of retention status.[35] $\Phi_{t,r}$ 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 unobservable "abilities." Hence the effect of treatment is both heterogeneous and time-varying.

The decision to retain a child is the solution to some complicated game being played between the parents, the teachers, the child and the school. While in principle we can think of modelling such a game, we choose to instead approximate it with a threshold crossing model as described in Section 2. As shown in Heckman and Navarro (2007), this model is in fact nonparametrically identified using the same arguments as in Section 3.1.

The actual form of the model for retention we use is the following.[36] We write the latent index V as

$$V_i\left(r\right) = \lambda_{0,r} + X_{i,r}\lambda_{x,r} + Z_{i,r}\lambda_{z,r} + A_i\rho_{A,r} + B_i\rho_{B,r} + C_i\rho_{C,r} + \sum_{\tau=2}^{r}\eta_i^{(\tau)}\psi_r^{(\tau)} + \upsilon_{i,r} \text{ for } r = \underline{R}, ..., \bar{R}.$$

$D_i\left(r\right)$ would then be defined as

$$D_i\left(r\right) = \mathbf{1}\left(V_i\left(r\right) > 0 | \{V_i\left(h\right) \leq 0\}_{h=1}^{r-1}\right).$$

Notice that, consistent with our data, we allow for exclusions in the index, so that some variables (Z) are included in the retention equations but not in the outcomes. In the data this corresponds to the 7 binary measures of the retention policies summarized in Table 2.[37] As discussed in Section 3.1, given that test scores in kindergarten are free of selection,

[35] The main reason we do this is to save on the number of parameters we are estimating. Preliminary reduced form regressions suggested that the marginal effects did not vary much across grades.

[36] 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. Given that the number of parameters we are estimating is already 616, and the number of parameters would increase considerably, we stick with the linear form.

[37] We examine whether these are valid exclusions in a simple two stage least squares regression and find

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. Furthermore, to address non-random sample attrition, we estimate 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 non-parametrically identified, as shown in Section 3.1. 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 multi-step estimation procedure, we estimate all of the parameters in the model jointly by maximum likelihood in a single step.

4.2 Results

In Online Appendix Tables D1 and D2 we present evidence of the fit of the model. We show that the model fits the means and variances of all the test measures very well, and 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. We cannot reject the hypothesis of equality of predicted and actual probabilities.[38]

Figure 1 presents evidence of selection on the different components of ability. Not surprisingly, students who are not retained have higher general and cognitive ability than those who are retained. In general, students who are retained early have lower ability than students who are retained in other grades. Kindergarten retainees generally have higher ability than early retainees but lower ability than late retainees. Interestingly, the distribution of behavioral ability is somewhat comparable for late retainees and students who are not retained in terms of the lower tail, but the upper tail of behavioral ability is actually higher for late-retainees.

These patterns can be understood by considering that a) the kindergarten retention decision is not as closely related with the abilities of the child, and b) dynamic selection. Because of a), early retention is more closely related to the abilities of the child. Hence, early retainees have lower abilities than kindergarten retainees. Then, because of b), late retainess have higher abilities than early retainees since the worse students have already been retained.

To place our estimates in context, Table 4 compares estimates of average treatment effects in reading scores (Panel A) and math scores (Panel B) using OLS, fixed effects and our factor method. The model is estimated jointly in each case, allowing a separate effect of

that they satisfy the test of overidentifying restrictions in this setting.

[38] Parameter estimates and standard errors are available in Online Appendix D.

retention in different years. For OLS, the math scores are used to control for selection (or unobservable "ability") in the reading equation and reading scores control for selection in the math equation. In what follows, we focus the discussion on results for reading in Panel A, though similar observations apply to math.

The initial effect of kindergarten retention on reading in 1999/00 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 2001/02 (column 3) the results become qualitatively different across the methods. OLS predicts that achievement decreases by 7% for students retained in kindergarten, whereas our model predicts that it increases by 4%. The fixed effect 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 effect and 0 for our model. Notice that OLS only controls for unobservable abilities in one dimension and through contemporaneous test scores in the other subject, whereas the measure of ability in our model takes into account the whole history of test scores, as well as controlling for different dimensions of ability. Particularly given the changing importance of different components of ability over time (as evidenced in the variance decomposition in Tables D3 and D4), it is not surprising that the results differ more across the two models as time passes. Furthermore, the fixed effect model, in assuming that ability is fixed and one-dimensional, cannot capture changing patterns of selection across different components of ability over time.

By 2003/04, 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 2001/02. In contrast, the fixed effect 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 the positive average treatment effects we estimate are not unique to our model.

Table 4 only compares estimates of *average* treatment effects across the OLS, fixed effect and our estimator. A key contribution of our estimator is to provide a method for estimating heterogeneous treatment effects that vary by unobservable student abilities. The OLS and fixed effect estimators are poorly equipped for such comparisons. For instance, if the retained students make smaller gains in the absence of retention (essential heterogeneity), the average treatment could overstate the benefits to retaining those students.

With this observation in mind, Table 5 also describes treatment on the treated (and the untreated) parameters for both reading and math test scores (Panels A and B respectively) in the 2003-04 school year. The predicted levels of achievement from which these gains are calculated are included in Online Appendix Table D5. The columns correspond to 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 (as reported in the last column of Table 4) for comparison.

Considering first the treatment on the treated parameters, students who are actually retained in kindergarten perform 6% lower in reading and math by 2003-04 than if they had not been retained. 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.

Thus, the treatment on the treated parameters suggest that the effect of retention is generally negative, in contrast to the average treatment effects reported in the last column (and in Table 4), which predict that the effect of retention in kindergarten is small or 0 and positive for early retention. Again, the effect is not statistically significantly different from 0 for late retention. We can see that these non-negative average treatment effects are driven by the untreated students. The finding that students who are not retained would actually be better off from retention than students who actually are retained is important. We provide some intuition behind this finding below.

4.2.1 Heterogeneity in Treatment Effects by Abilities

An advantage of our method is that we can provide new insight into the differences between the average treatment effect and the treatment on treated parameters by describing how treatment effects vary by the unobservable abilities of the student. Figure 2 shows how the treatment effect of being retained at different grades varies across the percentiles of the general, behavioral and cognitive ability distributions for reading and math by 2003/04. Comparing across graphs, we see that generally lower ability students experience losses (or are no better off) due to retention whereas the higher ability students benefit from retention. Thus, what the main pattern shows is that a high ability student would actually perform better by 2003/04 if retained relative to not being retained and receiving an additional year of course material. The opposite is true for low ability students. While at first these results may be surprising, they are intuitively appealing for several reasons.

First, the test scores reported in the ECLS-K are not actually those used to determine retention decisions. Thus, while we recognize the student as high ability from the history of their performance on these standardized exams, his performance in the classroom could suggest otherwise. This is further supported by the observation that, even if we restrict the sample to students whose achievement is below the median, this sample does not capture all retainees. Second, we permit the factor loadings to vary based on retention status. As shown in Online Appendix Table D10, generally the factor loadings are larger for the retained than for the not retained outcomes and positive in cognitive and general ability. 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.

High ability students 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.[39]

Additionally, it may be that teachers and/or parents put more resources into students who are retained. If high ability students are better-equipped to take advantage of these additional resources than low ability students, this may explain the difference across ability types. Furthermore, high ability students may have higher ability parents (assuming intergenerational transmission of human capital). These parents may be better-equipped to ensure that when their child is retained he gets the best teachers and the attention (and resources) he needs. Thus, resources may be invested disproportionately more in high ability students who are retained than in low ability students. Finally, high ability students who are retained may attend better schools, further reinforcing our argument.

4.2.2 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 2003/04 outcomes. For instance, for the case of late retention, the results reported in Table 5 and Figure 2 are short run effects, achievement gains 1 to 2 years after retention. For kindergarten retention, the effects are longer run, i.e., 4 to 5 years after treatment.

To consider how treatment effects vary over time, Figures 3 and 4 compare treatment effects of kindergarten and early retention at the different periods we observe in the data.

[39] This finding is further supported by research by Bedard and Dhuey (2006) and others suggesting that the age relative to other children in the classroom matters for performance.

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 respectively.[40] Figure 3 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 (2001) the average treatment effect is somewhat positive at 4%, and 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 3 shows a similar pattern for the treatment on the treated, i.e., students who are actually retained in kindergarten. The initial effect of retention is slightly more negative than the sample on average, -28% in reading and -19% in math. However, 2 years later the students have made significant progress and only perform about 9% lower in reading and 7% lower in math. The treatment on the treated remains negative in 2003/04 at about -6%. Thus there is some evidence that students catch up with where their achievement would be if not retained, though the rate of convergence diminishes over time.

With early retention, 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. Furthermore, the initial effect of early retention for early retainees is worse in reading than in math, -15% and -7% respectively. Notice that the initial shock for early retention is much smaller than for kindergarten. This could potentially be explained by the fact that early retainees can be retained in either first or second grade, so their initial effect may be up to 2 years after retention occurred. As in kindergarten, there is evidence that reading scores catch up over time. This does not appear to be the case for math.

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 4.2.1. Online Appendix Figures D1 and D2 show that, as before, this patterns follow because higher ability students generally fare better than low ability students when retained.

Overall evidence from considering the time elapsed since treatment suggests that students begin to recover from the initial negative shock from retention 2 years later (with the exception of early retainees in math). There is also evidence that the gains may level off over time, with the treatment effects remaining negative for the treated in our sample period.

[40] Online Appendix Tables D8 and D9 show the gains and standard errors for different time periods and correspond to the different points in these figures.

Interestingly, these findings contrast to evidence in the literature which suggests that any gain in retention may actually be short-lived.[41] Given that the initial effect of late retention is approximately 0 for reading and math, it could be the case that students in later grades actually experience long run gains, if similar patterns hold.

4.2.3 Marginal Policy Change

The evidence presented so far shows that the individuals who are actually retained are very different from the average individual. This is reflected in the fact that average treatment effects are significantly different than their treatment on the treated counterparts. The average treatment effect is helpful when evaluating a universal policy, while treatment on the treated is relevant for evaluating the effect of the retention policies already in place. However neither tells us, necessarily, about the effect of a marginal change in the retention policy.

As Figure 2 shows, there is considerable heterogeneity in treatment effects by abilities. Hence, 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.

We consider the effect of a marginal change in retention policies in Table 6. In particular, 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. We present three sets of results. In column 3, we show the gains in achievement for those students who are no longer retained as a consequence of the policy change. For comparison, column 4 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 5), while column 5 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).

For example, the first row of 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 3, we see that these marginal students gain 3% in both and math from the change in retention status to not being retained. In contrast, the average student who is not retained would lose 3% in reading and 1% in math by not being retained relative to being retained in kindergarten. The average student already being retained in kindergarten would gain 6% in reading and in math if he were not retained. Except for the case involving late retention in reading, where the estimate is very imprecise,

[41] See Frederick and Hauser (2006) for a summary of the literature.

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 is to be expected since there is a 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. Given the general positive relationship between ability and the benefits of retention described above, the marginal students will not benefit as much from not being retained as the average student who is already retained (i.e., the marginal students are not hurt as much by retention).

5 Conclusion

In this paper, we develop and apply a framework for the analysis of multiple treatment effects, focusing on the case of time-varying treatments. In our model, each treatment is associated with a treatment time. The main challenge is to distinguish the effects of treatment at different times from selection of different unobservable types into treatment at different times. Our method accounts for essential heterogeneity, i.e., that the gains to treatment vary by unobservable types and are taken into account in the selection decision. The additional challenge in the dynamic context is that selection into each treatment time is sequential. As a result of dynamic selection, existing methods that estimate static binary treatment effects cannot be easily extended to estimate time-varying treatment effects.

Our analysis of grade retention shows the importance of extending the standard static framework to estimate time-varying treatment effects. First, we find evidence of dynamic selection, which is not accounted for in previous studies in the literature. In particular, students who are retained in first or second grade have lower ability, in several dimensions, than students who are retained in kindergarten or third/fourth grade.

We also find that the effect of repeating a grade on tests scores varies considerably by student type, by the time at which the student is retained and by time elapsed since retention. In general, we find that the effect of retention is large and negative in the short run and that this effect diminishes (or even becomes positive) as time since retention passes. The effects tend to be more negative for the students being retained (treated students) than for the average student. Thus, estimates for the average student would not be very useful for policy.

The disparity between the treatment on the treated and treatment effect for the average student is because of unobserved ability. A key contribution of our approach is that it allows us to recover the distribution of the unobservables nonparametrically. Thus, we can show directly how the treatment effects vary by the abilities of the students. We find that the losses for retention are larger for low ability students. In fact, high ability students can even benefit from being retained in some cases. Overall, these results suggest that grade retention does not improve the performance of low achieving students.

Our findings also help illustrate the potential limitations of applying static methods to estimate time-varying treatment effects. Regression discontinuity designs can be a useful approach for estimating the effect of retention at a given grade. However, when there is a higher threshold for students to be promoted to the next grade, higher ability students will be retained. A regression discontinuity design that focuses on students close to this threshold may find a positive effect of retention, even if lower ability students are being hurt by the policy. Furthermore, if there is dynamic selection, comparing these policies across grades may not be straightforward, as the students at the margin of being retained are likely to differ across grades.

Our findings also suggest that differences in *the* estimated effect of retention across studies (see Holmes, 1989 and Jimerson, 2001) that focus on different grades may not be surprising. One source of these disparities is simply that different types of students are retained at different grades. A second reason is that, even after controlling for dynamic selection, we find that the effect of retention varies across grades.

Many policy evaluation problems are dynamic and would face similar challenges as those we have highlighted in our application. The method we develop can be applied to identify causal treatment effects in many other settings where heterogeneity in the effect of treatment across time and unobservables is likely to be important.

References

Abbring, J. H., and G. J. Van den Berg (2003): "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71(5), 1491–1517.

Angrist, J. D., and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430), 431–442.

Arcidiacono, P., J. Cooley, and A. Hussey (2008): "The Economic Return to an MBA," *International Economic Review*, 49(3), 873 – 899.

Bedard, K., and E. Dhuey (2006): "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects," *The Quarterly Journal of Economics*, 121(4), 1437–1472.

Billingsley, P. (1995): *Probability and measure*, A Wiley-Interscience publication. Wiley, New York, 3. ed edn.

Bonhomme, S., and J.-M. Robin (2010): "Generalized Non-parametric Deconvolution with an Application to Earnings Dynamics," *Review of Economic Studies*, 77(2), 491–533.

Carneiro, P., K. Hansen, and J. 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," *International Economic Review*, 44(2), 361–422, 2001 Lawrence R. Klein Lecture.

Cataneo, M. D. (2009): "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability," *Journal of Econometrics*, In Press, Corrected Proof.

Cellini, S. R., F. Ferreira, and J. Rothstein (2010): "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design," *Quarterly Journal of Economics*, 125(1), 215–261.

Cunha, F., J. Heckman, and S. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," NBER Working Papers 15664, NBER.

Cunha, F., J. J. Heckman, and S. Navarro (2005): "Separating Uncertainty from Heterogeneity in Life Cycle Earnings, The 2004 Hicks Lecture," *Oxford Economic Papers*, 57(2), 191–261.

——— (2007): "The Identification and Economic Content of Ordered Choice Models with Stochastic Cutoffs," *International Economic Review*, 48(4), 1273 – 1309.

Frederick, C. B., and R. M. Hauser (2006): "Have We Put an End to Social Promotion? Changes in Grade Retention Rates among Children Aged 6 to 17 from 1972 to 2003," Unpublished manuscript, University of Wisconsin-Madison.

Frölich, M. (2004): "Programme Evaluation with Multiple Treatments," *Journal of Economic Surveys*, 18(2), 181–224.

Gill, R. D., and J. M. Robins (2001): "Causal Inference for Complex Longitudinal Data: The Continuous Case," *The Annals of Statistics*, 29(6), 1785–1811.

Hahn, J., P. E. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

Ham, J. C., and R. J. LaLonde (1996): "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training," *Econometrica*, 64(1), 175–205.

Heckman, J. J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80(2), 313–318.

Heckman, J. J., V. J. Hotz, and J. R. Walker (1985): "New Evidence on the Timing and Spacing of Births," *American Economic Review*, 75(2), 179–184, Papers and Proceedings of the Ninety-Seventh Annual Meeting of the American Economic Association.

Heckman, J. J., and S. Navarro (2004): "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics*, 86(1), 30–57.

——— (2007): "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics*, 136(2), 341–396.

Heckman, J. J., and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman, and B. Singer, vol. 10, pp. 156–245. Cambridge University Press, New York.

Heckman, J. J., and J. A. Smith (1998): "Evaluating the Welfare State," in *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, ed. by S. Strom, pp. 241–318. Cambridge University Press, New York.

HECKMAN, J. J., S. URZUA, AND E. J. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

HECKMAN, J. J., AND E. J. VYTLACIL (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments," in *Handbook of Econometrics, Volume 6*, ed. by J. Heckman, and E. Leamer. Elsevier, Amsterdam, Forthcoming.

HECKMAN, J. J., AND J. R. WALKER (1990): "The Relationship Between Wages and Income and the Timing and Spacing of Births: Evidence from Swedish Longitudinal Data," *Econometrica*, 58(6), 1411–1441.

HOLMES, C. T. (1989): "Grade-level retention effects: A meta-analysis of research studies," in *Flunking grades: Research and policies on retention*, ed. by L. Shepard, and M. Smith, pp. 16–33. The Falmer Press, London.

HU, Y., AND S. M. SCHENNACH (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76(1), 195–216.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

JACOB, B. A., AND L. LEFGREN (2004): "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics*, 86(1), 226–244.

——— (2009): "The Effect of Grade Retention on High School Completion," *American Economic Journal: Applied Economics*, 1(3), 33–58.

JIMERSON, S. R. (2001): "Meta-analysis of grade retention research: Implications for practice in the 21st century," *School Psychology Review*, 30(3), 420–437.

JÖRESKOG, K. G., AND A. 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(351), 631–639.

KOTLARSKI, I. I. (1967): "On Characterizing the Gamma and Normal Distribution," *Pacific Journal of Mathematics*, 20, 69–76.

LECHNER, M. (2004): "Sequential Matching Estimation of Dynamic Causal Models," Discussion Paper 2004, IZA Discussion Paper.

Matzkin, R. L. (2003): “Nonparametric Estimation of Nonadditive Random Functions,” *Econometrica*, 71(5), 1339–1375.

Murphy, S. A. (2003): “Optimal Dynamic Treatment Regimes,” *Journal of the Royal Statistical Society, Series B*, 65(2), 331–366.

Nagaoka, J., and M. Roderick (2005): “Retention Under Chicago’s High-Stakes Testing Program: Helpful, Harmful, or Harmless?,” *Educational Evaluation and Policy Analysis*, 27(4), 309–340.

Navarro, S. (2008): “Control Function,” in *The New Palgrave Dictionary of Economics.*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan Press, London, second edn.

Nekipelov, D. (2008): “Endogenous Multi-Valued Treatment Effect Model under Monotonicity,” Unpublished manuscript, Berkeley.

Prakasa Rao, B. (1992): *Identifiability in Stochastic Models: Characterization of Probability Distributions*, Probability and mathematical statistics. Academic Press, Boston.

Schennach, S. M. (2004): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica*, 72(1), 33–75.

A Tables

Table 1: Summary Statistics

	Value of Variables in 1998-99 School Year for Observations Included in:					
	1998-99 School Year			2003-04 School Year		
Variables	Observation	Mean	Standard Deviation	Observation	Mean	Standard Deviation
General Test Score	7549	3.09	0.35	2078	3.14	0.33
Reading Test Score	7608	3.36	0.28	2078	3.39	0.27
Math Test Score	7794	3.10	0.36	2101	3.14	0.35
Approach to Learning	7829	0.05	0.98	2104	0.13	0.95
Self-Control	7808	0.03	0.97	2097	0.11	0.94
Interpersonal Skills	7782	0.02	0.98	2095	0.09	0.96
Male	7832	0.50	0.50	2106	0.49	0.50
White	7832	0.65	0.48	2106	0.77	0.42
Black	7832	0.12	0.32	2106	0.07	0.26
Hispanic	7832	0.14	0.34	2106	0.09	0.28
Body Mass Index	7832	16.25	2.13	2106	16.21	2.10
Age	7832	5.62	0.34	2106	5.63	0.34
Number of Siblings	7832	1.42	1.11	2106	1.41	1.07
Socioeconomic Status Index	7832	0.10	0.78	2106	0.20	0.74
Attended Full Time Kindergarten	7832	0.58	0.49	2106	0.52	0.50
TV Rule at Home	7832	0.89	0.32	2106	0.89	0.31
Mother in Household	7832	0.01	0.11	2106	0.01	0.11
Father in Household	7832	0.17	0.37	2106	0.12	0.32
Number of Books at home	7832	80.54	60.75	2106	88.76	60.23
Minority Students in School between (1%,5%)	7832	0.20	0.40	2106	0.20	0.40
Minority Students in School between (5%,10%)	7832	0.15	0.36	2106	0.12	0.33
Minority Students in School between (10%,25%)	7832	0.10	0.30	2106	0.05	0.22
Minority Students in School >25%	7832	0.16	0.36	2106	0.09	0.29
Public School	7832	0.78	0.42	2106	0.73	0.44
TT1 Funds Received by School	7832	0.62	0.49	2106	0.63	0.48
Crime a Problem	7832	0.46	0.58	2106	0.36	0.52
Students Bring Weapons	7832	0.16	0.37	2106	0.13	0.34
Children or Teachers Physically Attacked	7832	0.36	0.48	2106	0.35	0.48
Security Measures in School	7832	0.55	0.50	2106	0.58	0.49
Parents Involved in School Activities	7832	2.97	0.90	2106	3.10	0.83
Teacher has a Master's Degree	7832	0.35	0.48	2106	0.34	0.48
Teacher Experience	7832	14.31	9.03	2106	14.39	8.97
Student's Class Size	7832	20.40	5.00	2106	19.89	4.80
Teacher's Rating of Class Behavior	7832	1.56	0.78	2106	1.52	0.77
Minority Students in Class between (1%,5%)	7832	0.08	0.26	2106	0.09	0.29
Minority Students in Class between (5%,10%)	7832	0.13	0.33	2106	0.16	0.36
Minority Students in Class between (10%,25%)	7832	0.18	0.39	2106	0.18	0.38
Minority Students in Class >25%	7832	0.42	0.49	2106	0.28	0.45

Source: ECLS-K Longitudinal Kindergarten-Fifth Grade Public-Use Data File

Note: For our counter-factual 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 2003-04 school year.

Table 2: Summary Statistics for Selected Variables by Retention Status (1998/1999 School Year)

	Not Retained		Retained in Kindergarten		Retained Early		Retained Late	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
General Test Score	3.12	0.33	2.85	0.37	2.72	0.33	2.78	0.32
Reading Test Score	3.39	0.27	3.13	0.21	3.08	0.18	3.15	0.17
Math Test Score	3.14	0.35	2.77	0.32	2.67	0.26	2.74	0.25
Approach to Learning	0.12	0.94	-0.72	0.99	-0.91	0.95	-0.40	0.98
Self-Control	0.06	0.96	-0.31	1.02	-0.41	1.03	-0.09	0.93
Interpersonal Skills	0.06	0.96	-0.36	0.95	-0.53	1.00	-0.21	1.01
Male	0.49	0.50	0.66	0.48	0.63	0.48	0.54	0.50
Black	0.11	0.31	0.14	0.35	0.29	0.46	0.28	0.45
Hispanic	0.13	0.34	0.12	0.32	0.19	0.39	0.18	0.39
Age	5.64	0.34	5.39	0.28	5.50	0.32	5.52	0.33
Attended Full Time Kindergarten	0.57	0.49	0.62	0.49	0.61	0.49	0.72	0.45
Number of Siblings	1.39	1.08	1.65	1.27	1.80	1.41	1.52	1.25
Socioeconomic Status Index	0.13	0.77	-0.12	0.80	-0.33	0.69	-0.54	0.60
TV Rule at Home	0.89	0.31	0.90	0.30	0.83	0.37	0.90	0.31
Father in Household	0.16	0.37	0.19	0.39	0.28	0.45	0.38	0.49
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%	0.15	0.36	0.16	0.37	0.27	0.44	0.38	0.49
Public School	0.77	0.42	0.73	0.44	0.91	0.28	0.93	0.25
TT1 Funds Received by School	0.62	0.49	0.61	0.49	0.76	0.43	0.79	0.41
Teacher has a Master's Degree	0.35	0.48	0.32	0.47	0.40	0.49	0.33	0.47
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%	0.40	0.49	0.42	0.50	0.63	0.48	0.66	0.48
Policy: Can be Retained for Immaturity	0.76	0.43	0.78	0.41	0.72	0.45	0.68	0.47
Policy: Can be Retained at Parents Request	0.75	0.43	0.76	0.43	0.79	0.41	0.76	0.43
Policy: Can be Retained due to Academic Deficiencies	0.88	0.33	0.83	0.38	0.91	0.29	0.88	0.32
Policy: Can be Retained Any Grade More than Once	0.10	0.30	0.13	0.33	0.14	0.35	0.15	0.36
Policy: Can be Retained More than Once in Elementary School	0.35	0.48	0.30	0.46	0.43	0.50	0.50	0.50
Policy: Can be Retained Without Parents Permission	0.44	0.50	0.45	0.50	0.61	0.49	0.58	0.50
Observations	7038		255		288		87	

Source: ECLS-K Longitudinal Kindergarten-Fifth Grade Public-Use Data File

Note: For our counter-factual 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 2003-04 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 necessarilly 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 (7832). This is because we don't know every childrens' 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

Panel A: Reading Score

Dependent Variable	Kindergarten Reading Score[#]	Reading Score for 2003-04 School Year		
Retained in Kindergarten	-0.1775*	-0.1791*	-0.0948*	-0.0926*
Retained Early (1st or 2nd grade)	-0.2014*	-0.2306*	-0.1450*	-0.1374*
Retained Late (3rd or 4th grade)	-0.1222*	-0.1192*	-0.0498	-0.0358
Student's Characteristics	Yes	Yes	Yes	Yes
Family Characteristics	Yes	Yes	Yes	Yes
School Characteristics	Yes	Yes	Yes	Yes
Age and Age Squared	Yes	Yes	Yes	Yes
Kindergarten Cognitive Tests	--	No	Yes	Yes
Kindergarten Behavioral Ratings	--	No	No	Yes
No. of Observations	5319	2040	2014	1998
P-value for KI = EA = LA[+]	0.003	0.019	0.026	0.012
P-value for KI = EA	0.189	0.099	0.079	0.113
P-value for EA = LA	0.001	0.006	0.009	0.003
P-value for KI = LA	0.028	0.148	0.192	0.092
R squared	0.312	0.385	0.530	0.530

Panel B: Math Score

Dependent Variable	Kindergarten Reading Score[#]	Reading Score for 2003-04 School Year		
Retained in Kindergarten	-0.2735*	-0.1804*	-0.0727*	-0.0889*
Retained Early (1st or 2nd grade)	-0.3172*	-0.2450*	-0.1463*	-0.1396*
Retained Late (3rd or 4th grade)	-0.2240*	-0.1697*	-0.0875*	-0.0387
Student's Characteristics	Yes	Yes	Yes	Yes
Family Characteristics	Yes	Yes	Yes	Yes
School Characteristics	Yes	Yes	Yes	Yes
Age and Age Squared	Yes	Yes	Yes	Yes
Kindergarten Cognitive Tests	--	No	Yes	Yes
Kindergarten Behavioral Ratings	--	No	No	Yes
No. of Observations	5462	2043	2017	1998
P-value for KI = EA = LA[+]	0.006	0.094	0.086	0.012
P-value for KI = EA	0.097	0.071	0.038	0.076
P-value for EA = LA	0.002	0.079	0.097	0.004
P-value for KI = LA	0.136	0.813	0.684	0.141
R squared	0.408	0.357	0.531	0.522

* Statistically significant at 5% level

[#] 1998-99 School Year

[+] KI, EA, and LA stand for the coefficient of the dummy variable for "retained in kindergarten", "retained early", and "retained late",

Note: P values less than 0.05 are shaded, and indicates rejection of the hypothesis of equality at the 5% confidence level. Yes/No indicates if each group of variables is included as controls.

Table 4: Estimated Coefficients for Retention Variables in Outcome Equation

Panel A: Reading Score

		Outcome Equation in 1999-2000 School Year	Outcome Equation in 2001-02 School Year	Outcome Equation in 2003-04 School Year
Retained in Kindergarten	OLS	-0.241	-0.068	-0.065
	Fixed Effect	-0.283	-0.008	0.051
	Model	-0.263	0.041	0.025
Retained Early	OLS	--	-0.146	-0.080
	Fixed Effect	--	-0.049	0.062
	Model	--	0.004	0.046
Retained Late	OLS	--	--	0.014
	Fixed Effect	--	--	0.074
	Model	--	--	0.056

Panel B: Math Score

		Outcome Equation in 1999-2000 School Year	Outcome Equation in 2001-02 School Year	Outcome Equation in 2003-04 School Year
Retained in Kindergarten	OLS	-0.025	-0.050	-0.049
	Fixed Effect	-0.099	0.071	0.151
	Model	-0.117	0.039	0.004
Retained Early	OLS	--	-0.040	-0.060
	Fixed Effect	--	0.039	0.116
	Model	--	-0.053	0.066
Retained Late	OLS	--	--	-0.091
	Fixed Effect	--	--	0.075
	Model	--	--	0.083

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).

Table 5: Average Test Score Gain by Retention Status: 2003-04 School Year

Panel A: Reading Score

Average Gain	A student who is actually (i.e. conditional on the retention status being:)				ATE (unconditional)
	Not Retained	Retained in Kindergarten	Retained Early	Retained Late	
Retained in Kindergarten vs Not Retained	0.034 (0.014)	-0.057 (0.013)	-0.086 (0.018)	-0.023 (0.027)	0.025 (0.012)
Retained Early vs Not Retained	0.058 (0.019)	-0.092 (0.019)	-0.111 (0.023)	-0.046 (0.046)	0.046 (0.017)
Retained Late vs Not Retained	0.058 (0.112)	0.026 (0.058)	0.016 (0.080)	0.022 (0.084)	0.056 (0.101)

Panel B: Math Score

Average Gain	A student who is actually (i.e. conditional on the retention status being:)				ATE (unconditional)
	Not Retained	Retained in Kindergarten	Retained Early	Retained Late	
Retained in Kindergarten vs Not Retained	0.011 (0.024)	-0.057 (0.019)	-0.084 (0.021)	-0.071 (0.031)	0.004 (0.022)
Retained Early vs Not Retained	0.079 (0.021)	-0.058 (0.015)	-0.095 (0.017)	-0.016 (0.036)	0.066 (0.019)
Retained Late vs Not Retained	0.098 (0.337)	-0.075 (0.142)	-0.112 (0.162)	-0.052 (0.258)	0.083 (0.309)

Note: Let R = 1,2, 3, or ∞ represent the actual retention status of a student: retained in kindergarten, retained early (at grade 1 or 2), or retained late (at grade 3 or 4), never retained, respectively. Let $\zeta(i)$ be the potential test score if the student were retained at time i=1,2,3,∞. 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 were retained at 1 or 2 grade instead. Bootstrap standard errors are in parentheses.

Table 6: Policy Simulation Treatment Parameters: 2003-04 School Year

Panel A: Reading Score

Retention Status		Average Test Score if Not Retained minus Test Score if Retained Conditional on:		
Old Policy	New Policy	Changing to Not Retained	Original Retention Status	Not Retained
Kindergarten	Not retained	0.032	0.057	-0.034
Early	Not retained	0.066	0.111	-0.058
Late	Not retained	-0.096	-0.022	-0.058

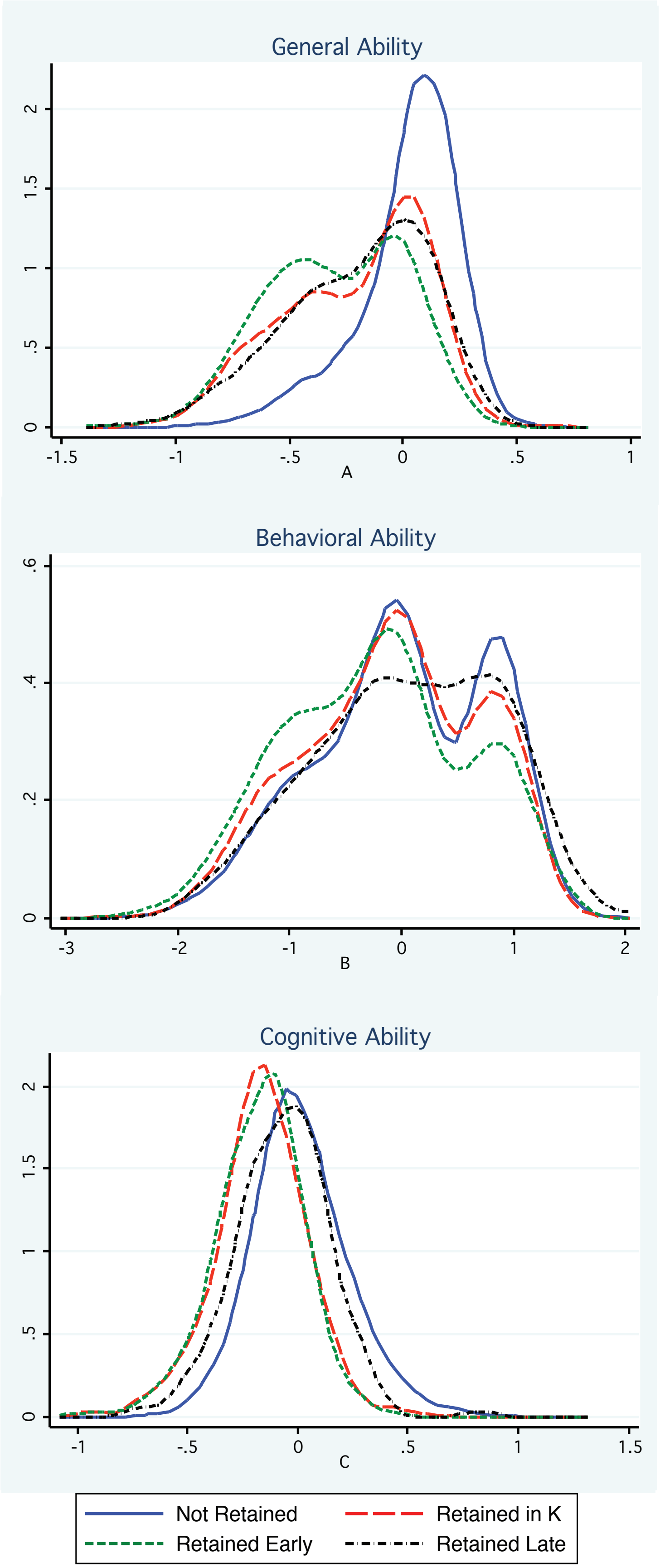
Panel B: Math Score

Retention Status		Average Test Score if Not Retained minus Test Score if Retained Conditional on:		
Old Policy	New Policy	Changing to Not Retained	Original Retention Status	Not Retained
Kindergarten	Not retained	0.029	0.057	-0.011
Early	Not retained	0.070	0.095	-0.079
Late	Not retained	-0.033	0.052	-0.098

Note: We fix all retention policy variables in Table 2 to 0 for all individuals. That is we make it harder for children 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 3 reports $E(\zeta_1\text{-}\zeta_0 \mid R_1 \neq R_0, R_1=\infty)$, column 4 reports $E(\zeta_1\text{-}\zeta_0 \mid R_0)$ and column 5 reports $E(\zeta_1\text{-}\zeta_0 \mid R_1=\infty)$. Notice that while some people 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.

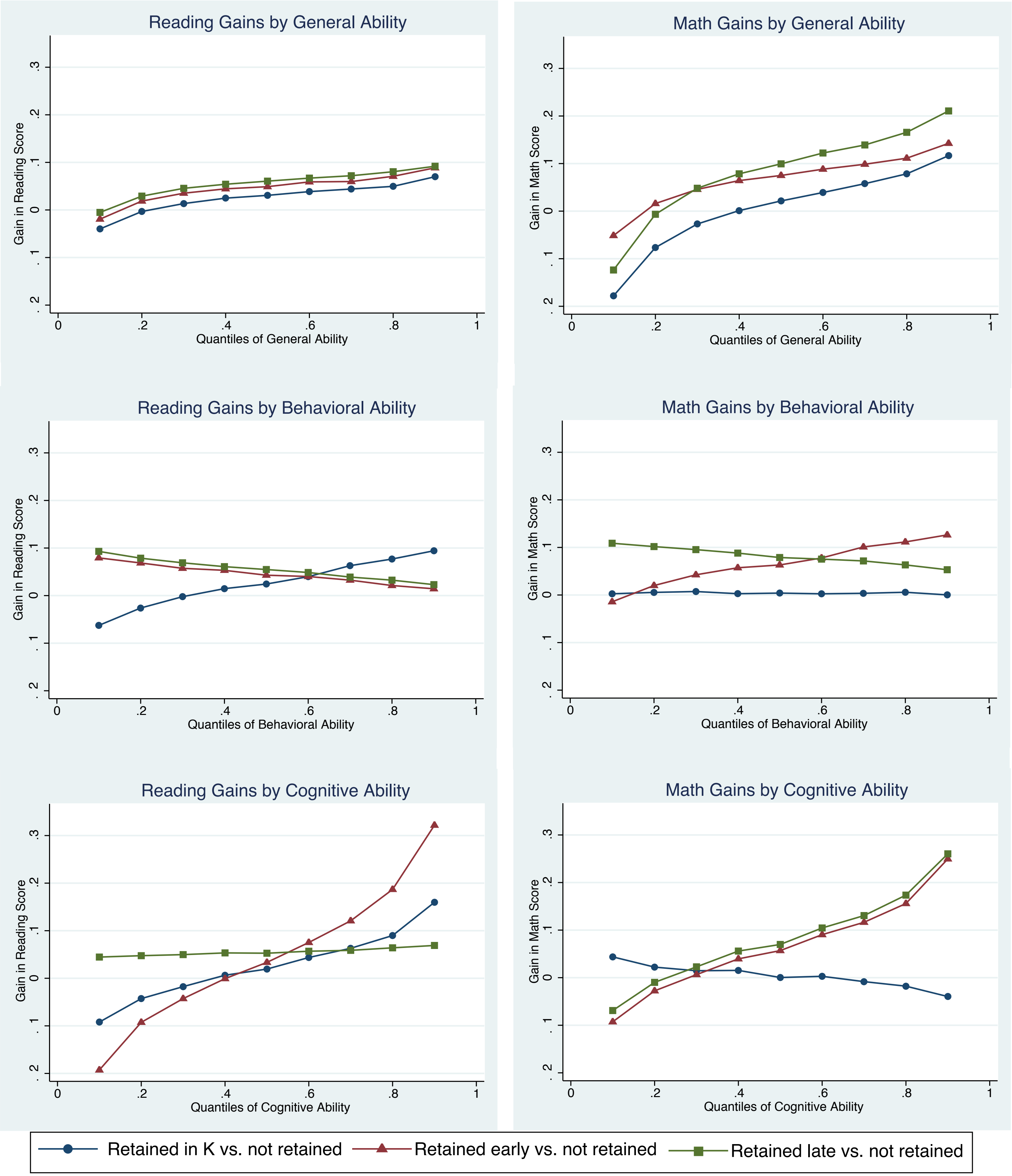
B Figures

Figure 1: Densities of Abilities by Retention Status



Note: Let f(X) denote the probability density function of ability X={A,B,C}. We allow f(X) to follow a mixtureof normals distribution. Let R={1,2,3,∞ } denote retention status: retained in kindergarten, retained early (1 or 2 grade), retained late (3 or 4) and not retained. The graph shows f(X|R=r) for each retention status.

Figure 2: Achievement Gains in 2003/04 by Ability Quantiles



Reading Gains by General Ability

Math Gains by General Ability

Reading Gains by Behavioral Ability

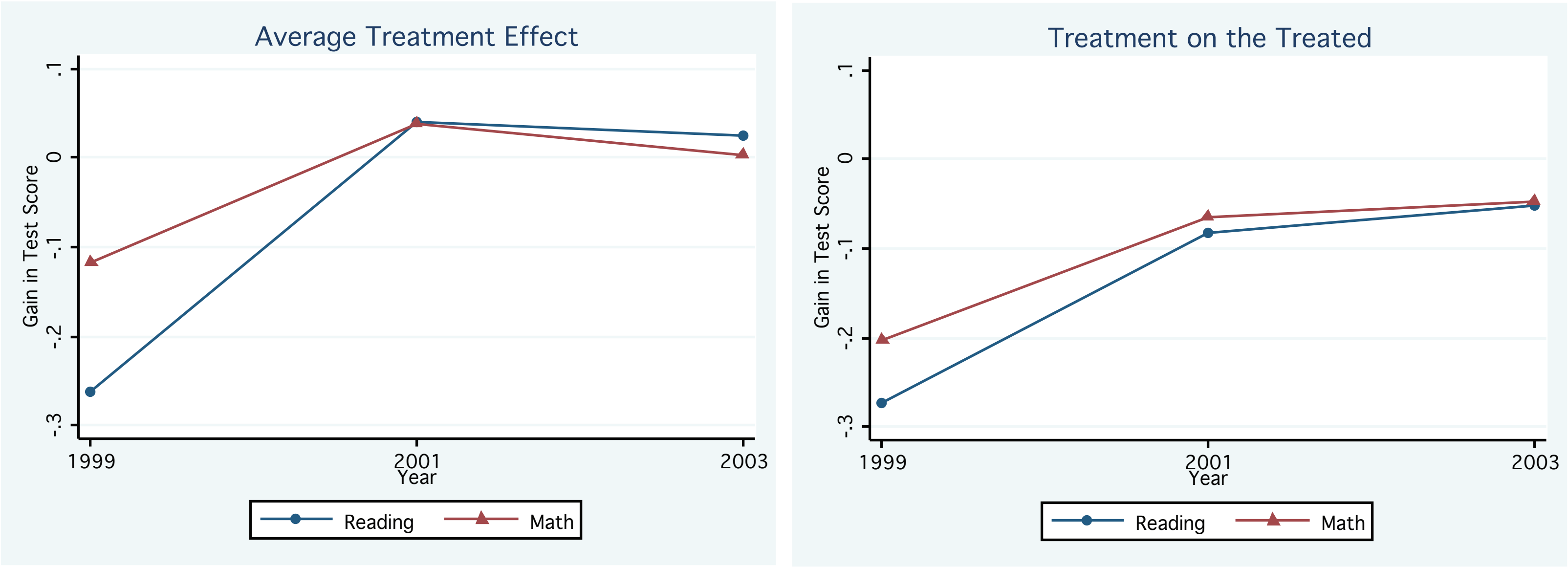
Math Gains by Behavioral Ability

Reading Gains by Cognitive Ability

Math Gains by Cognitive Ability

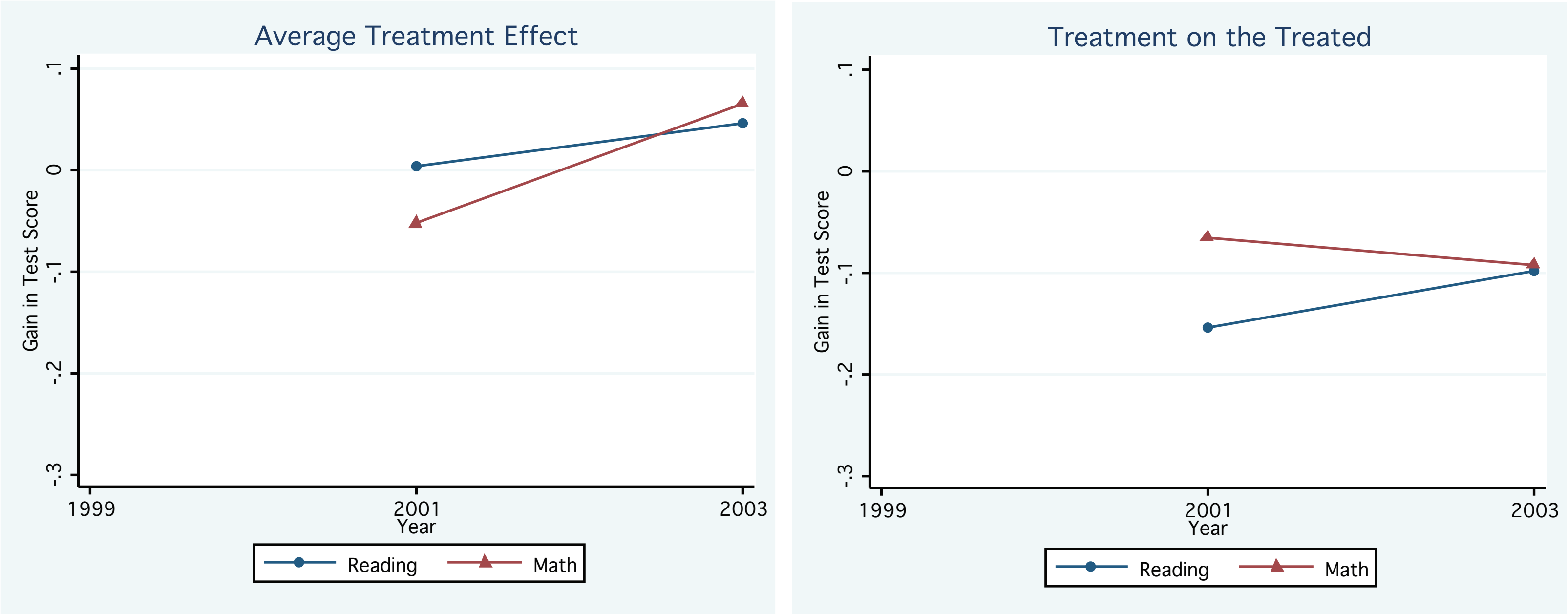
Note: Let $\zeta(t,r)$ and $\zeta(t,\infty)$ be the potential test scores at period t if the student is retained at r and if the student is not retained at all, respectively. Let X denote one kind of ability (i.e., etiher A,B or C). The graphs show $E[\zeta(t,r)-\zeta(t,\infty)|X=q]$ where q is the q^{th} quantile of the X-type of ability distribution.

Figure 3: Achievement Gains for Kindergarten Retention over Time



Note: Let $\zeta(t,1)$ and $\zeta(t,\infty)$ be the potential test scores at time t if the student is retained in kindergarten and if the kid is not retained at all, respectively. Let R={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 $E[\zeta(t,1)-\zeta(t,\infty)|R=t]$.

Figure 4: Achievement Gains for Early Retention over Time



Note: Let $\zeta(t,1)$ and $\zeta(t,\infty)$ be the potential test scores at time t if the student is retained in kindergarten and if the kid is not retained at all, respectively. Let R={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=2, and 3 for each test score. The Treatment on the Treated graph shows $E[\zeta(t,1)-\zeta(t,\infty)|R=t]$.