A Framework for the Analysis of Time-varying Treatment Effects: How The Timing of Grade Retention Affects Outcomes[*]

Preliminary and Incomplete.

Please do not cite or circulate.

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

October 29, 2009

Abstract

In this paper, we develop a method to estimate time-varying treatment effects in situations where dynamic selection into treatment may confound estimates of the treatment effect. Doing so, emphasizes an important policy tool, the timing of treatment. We illustrate the challenges in using conventional approaches to control for selection in a static setting for a dynamic setting. The method we develop assumes that the unobservables that jointly determine selection into treatment and the treatment effects can be modeled through a factor structure. We demonstrate how this method recovers a much richer set of counterfactuals than can be recovered with existing methods. We then apply our method to the study of grade retention using the Early Childhood Longitudinal Study of Kindergartners. We find evidence of dynamic selection and that the effect depends both on the time at which the student is retained and the time elapsed since 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 participants at the Applied Microeconomics lunch at UW-Madison, Anirban Basu, David Meltzer 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 are inherently dynamic in nature. In the case of the return to an advanced degree like an MBA (e.g. Arcidiacono et al. (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 et al. (1985), Heckman & 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. Furthermore these effects may differ depending on how much time has elapsed since treatment. 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.

In this paper, we develop a simple framework for the analysis of models with time-varying treatment effects. We focus on the single treatment case, 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 grade retention at different grades on student achievement at different times since retention. Grade retention is a controversial education policy and estimated effects on student achievement are mixed (e.g., Holmes 1989, Jimerson 2001, Jacob & Lefgren 2004). Yet, with the advent of nationwide accountability policies in the U.S., it is becoming increasingly common to tie grade promotion to performance on state standardized exams. We provide new evidence on the dynamic effects of grade retention, particularly informing the timing of retention decisions.

Like in the static framework, the key challenge is to distinguish the effect of treatment from selection of certain 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 that the treatment effect may vary over time. For example, students who are retained in kindergarten may differ in unobservable ways from students who are retained in second grade, leading to dynamic selection, which, in the absence of controls, is confounded with estimates of the effect of retention at a particular grade on achievement.

To motivate our preferred approach and to highlight the challenges with identifying time-varying treatment effects in the context of dynamic selection, we consider whether various methods that are applied to control for selection in the static context can be extended to our setting. We demonstrate that, because of the sequential nature of the decisions process (i.e., the dynamics) some potentially important counterfactual outcomes are simply lost with randomized control trials. That is, even in the potentially best case scenario of a repeated randomized experiment (or with some versions of repeated conditional independence, Lechner (2004)) many treatment on the treated type parameters cannot be recovered without invoking stronger assumptions. For instance, we cannot form the counterfactual expected achievement of being retained in period 1 for students who are retained in period 2 because in period 1 (when the randomization would need to occur), we do not yet know whether they will be retained in period 2. Thus, randomized control trials cannot tell us whether it would have been better to retain in grade 1 a student who is actually retained in grade 2, without strong assumptions on the heterogeneity.

We then consider the control function approach, which controls for unobserved heterogeneity through modelling the dependence of unobservables on observables via the selection equation (see Heckman & Navarro (2004)). For instance, the observed selection of students into being retained in different grade levels controls for unobserved student types in order to recover the true effect of retention on test scores.[1] Interestingly, we show that applying the control function approach results in the same set of missing counterfactuals as the repeated randomized control trial.

We also discuss the difficulty with applying widely-used instrumental variables methods to the dynamic setting. As in the static binary treatment case, when the unobservable (to the econometrician) gains to treatment are correlated with treatment selection ("essential heterogeneity" in the language of Heckman et al. (2006)), one of the more challenging aspects of this strategy is determining precisely what treatment effect is being estimated. While local methods have been designed to recover parameters of interest when essential heterogeneity is present, like local average treatment effects (Imbens & Angrist (1994)) or regression discontinuity designs (Hahn et al. (2001)), these types of parameters that compare individuals only at the margins are much harder to interpret when we add dynamics to the model. For example, the students who are on the margin of being retained in kindergarten may not be the marginal students in first grade.

Because of the challenges with extending these common existing methods to the dynamic setting, we develop a new approach. We explore a generalization of the semiparametric factor structure of Carneiro et al. (2003) in which it is assumed that a low dimensional set of unobservables affects all elements of the model. This strategy effectively places restrictions on the covariances between unobservables in the outcome and selection equations. For example, the same unobserved ability (or abilities) affects both the probability of being retained and test scores. While ability is time invariant, the factor structure we propose is quite general in that the effect of ability can vary both over time and based on treatment. Furthermore, given that we observe cognitive and behavioral measures, we can generalize to allow for different types of "ability" that might affect both the retention decision and the test scores.[2] We can also permit individual unobservable shocks to be correlated over time, such as multiple draws of a bad teacher.

The factor structure has several appealing aspects. First, it solves the problem of the missing counterfactuals and is arguably less data hungry than instrumental variable methods (or repeated randomized control trials). Second, because the unobserved components have an intuitive interpretation, we can determine whether students with lower general, cognitive or behavioral ability are more likely to be retained and how the treatment effect of retention varies by their ability type.

[1] Where we invoke similar assumptions as in the binary treatment frameworks where instrumental variables are available or certain curvature restrictions are satisfied such that the treatment choice is not collinear with the outcomes as in Heckman & Navarro (2007).

[2] As we show, the factor structure we propose maps directly to our empirical example and so it is very convenient but other normalizations are possible (Cunha et al. (2006), Robin & Bonhomme (2008)).

Third, because we are modeling the selection process, we can also provide insight into how to create better rules for when to retain students. In particular, if we think that retention decisions take into account observed test outcomes, we can obtain unbiased estimates of how test scores affect selection because we control for ability in the selection equation. This is not possible with other methods.

Our framework is similar to the one used for models with multiple potential treatments. The key difference is that in the dynamic case a treatment is equivalent to being treated at a given time, i.e., being retained in kindergarten or first grade. Secondly, because selection into treatment occurs sequentially over time the model is closer in spirit to an ordered model (generalized as in Cunha et al. (2007)). Standard multinomial and ordered choice models generally cannot be applied to our setting, unless the dynamic selection problem can be ignored and the sequence of participation is determined from the beginning.

Our approach to the study of dynamic treatments is close to that in Heckman & Navarro (2007) and similar in spirit to the analysis of Eberwein et al. (1997) and Abbring & Van den Berg (2003). As opposed to Eberwein et al. (1997) we do not focus on numerical problems and estimation but rather on identification. We share our interest on allowing for unobserved heterogeneity and dynamic selection. This is in contrast to the analyses of Gill & Robins (2001), Murphy (2003) and Lechner (2004), which are based on sequential conditional independence assumptions and so rule out selection on unobservables. Furthermore, our focus is not on the design of optimal treatment regimes as in Murphy (2003) but rather on the identification of treatment effects.

We apply our method to provide new evidence on the effect of retention on achievement using data from the Early Childhood Longitudinal Study of Kindergartners. Existing studies that deal convincingly with the selection problem and compare the effect of grade retention on student achievement across grades (such as Jacob & Lefgren (2004)) use a regression discontinuity framework, focusing on the treatment effects for the marginal students. Our study provides important new insight into how treatment effects vary by students of different abilities as well as by the timing and time elapsed since retention. We find evidence of dynamic selection and heterogeneous treatment effects by unobservable student abilities. We provide a much richer picture of the potential consequences of recent accountability policies that require that students meet a given threshold to be promoted to the next grade. For instance, an interesting implication of our results is that the higher the achievement threshold for promotion, the more likely that the policy will improve student achievement, as higher ability students benefit more for retention than lower ability students. This may have important implications for reconciling some of the mixed evidence regarding the effects of retention in the literature.[3]

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 show the problems associated with standard experimental and instrumental variable methods to recover counterfactuals in the dynamic heterogeneous case and ways to extend these methods. We then propose to impose a factor structure in the next section and prove it is semiparametrically identified. We discuss our application in Section

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

4.

2 The Framework

Consider the generic problem of evaluating the efficacy of a treatment. Let $P \in \mathcal{P} = \{1, 2, ..., \bar{P}\}$ index calendar time and $i = 1, ..., I$ index the individual. Since we allow for the treatment to be taken at different times, we define a random variable T that indicates the time at which treatment is received. We assume that treatment is taken at most once.[4] We let $T \in \mathcal{T} \subseteq \mathcal{P}$, that is treatment may be received at any time or only in a subset if treatment selection is limited to a subperiod so $\mathcal{T} = \{\underline{T}, \underline{T} + 1, ..., \bar{T} - 1, \bar{T}\}$ for $\underline{T} \geq 1$ and $\bar{T} \leq \bar{P}$. We adopt the convention of letting $T = 0$ for the "never" treated state.[5]

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

$$Y_i(P) = \sum_{\tau=\underline{T}}^{\bar{T}} D_i(\tau) \left[Y_i(P, \tau) - Y_i(P, 0)\right] + Y_i(P, 0) . \tag{1}$$

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 $\underline{T}$, the treated at time $\underline{T} + 1$, etc. Because of the sequential nature of the problem, by letting $Y_i(P, T)$ depend on treatment time, we allow for the possibility that the effect of treatment is dynamic in the sense that it does not only depend on receipt but the time treatment was received. It can also be interpreted as depending on the time since treatment $(P - T)$, making it straightforward to analyze the outcomes as durations, counts, etc.

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

A-1 *(No anticipations)* $Y_i(P, T) = Y_i(P, 0) = Y_i(P)$ *for* $T \geq P$.

That is, we rule out that potential outcomes that should be the same ex ante differ because *in the future* treatment times will be different. In our application, this means, for example, that after conditioning on all prior information, the fact that a student will be retained in second grade does not directly affect her performance in first grade. Importantly, this assumption should not be confused with the assumption that individuals are not forward looking, which the name may imply. Assumption **A-1** does not rule out that individuals may predict that they are more likely to get treated at a particular time T (i.e., have some anticipation as to treatment time).[6]

[4] 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 as shown in Appendix A.1.

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

[6] We follow the literature and refer to **A-1** as "no anticipations." What it rules out is that, after conditioning on

We further write the outcomes as

$$Y_i\left(P,T\right)=\Phi\left(P,T\right)+\epsilon_i\left(P,T\right), \tag{2}$$

where, because of **A-1**, we impose $\Phi\left(P,T\right)=0$ and $\epsilon_i\left(P,T\right)=\epsilon_i\left(P\right)$ if $T\geq P$.[7]

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 & Navarro (2007). If we define the treatment time specific index $V_i\left(T\right)=\lambda\left(T\right)+U_i\left(T\right)$,[8] then treatment time is selected according to

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

where $\mathbf{1}\left(a\right)$ 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 T can only be selected if treatment has not been taken before. We describe the generalization to the case where treatment can be received more than once in Appendix A.1.[9]

The observed outcome in period P will then be given by

$$Y_i\left(P\right)=\Phi(P,0)+\varepsilon(P,0)+\sum_{\tau=\underline{\mathrm{T}}}^{min\{P,\bar{T}\}}D_i\left(\tau\right)\left(\Phi\left(P,\tau\right)-\Phi(P,0)\right)+\sum_{\tau=\underline{\mathrm{T}}}^{min\{P,\bar{T}\}}D_i\left(\tau\right)\left(\epsilon_i\left(P,\tau\right)-\varepsilon\left(P,0\right)\right).$$

If there is no selection based on unobservables, then the problem becomes quite easy 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 to have higher test scores. Our goal in general will be to allow $\left(\epsilon_i\left(P,T\right),\epsilon_i\left(P',T''\right),U_i\left(T'''\right),U_i\left(T''''\right)\right)$ to be all correlated.

the information available at the pre-T period of interest P, the actual event of getting treated at time T has an effect on pre-time T 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 & Van den Berg (2003) and Heckman & Navarro (2007) for a discussion.

[7] 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 $Y_i^*\left(P,T\right)=\Phi\left(P,T\right)+\epsilon_i\left(P,T\right)$ so that the measured outcome $Y_i\left(P,T\right)$ would be $Y_i\left(P,T\right)=\mathbf{1}\left(Y_i^*\left(P,T\right)>0\right)$ where the function $\mathbf{1}\left(a\right)$ takes value 1 if a is true and 0 if it is not. Furthermore, additive separability in outcomes is not required, it can be relaxed using the analysis in Matzkin (2003).

[8] Additive separability is assumed for simplicity and can be relaxed using the analysis in Matzkin (1992).

[9] As we show in the Appendix, the extension to the case in which treatment can be received more than once requires a multiple spell model where the whole sequence of prior treatments/no treatments can potentially affect the decision each period. We can easily accommodate such a case at the cost of considerable notation where now T would be a vector containing the treatment history up to P and an individual would choose treatment every time the index becomes positive (not only the first time).

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

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

measures the effect at period P of receiving treatment at time T versus receiving treatment at time T'. If we let $T'=0$, this parameter would measure the effect at P of receiving treatment at time T versus not receiving treatment at all. An example of this first parameter would be the difference in earnings at age 30 for an individual if he repeats first grade versus if he repeats third grade.

The second individual parameter of interest

$$\begin{aligned}\Delta_i^2\left(\tau,T,T'\right) &= \left[Y_i\left(T+\tau,T\right)-Y_i\left(T+\tau,0\right)\right]-\left[Y_i\left(T'+\tau,T'\right)-Y_i\left(T+\tau,0\right)\right]\\ &= \Phi\left(T+\tau,T\right)-\Phi\left(T'+\tau,T'\right)+\epsilon_i\left(T+\tau,T\right)-\epsilon_i\left(T'+\tau,T'\right), \text{ for } \tau>0\end{aligned}$$

measures the difference in the effect of receiving treatment versus not receiving treatment τ periods after treatment time for two different treatment times T and T'. An example of this parameter would be the difference in test scores 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 dynamic 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 $T=t$ versus not receiving treatment

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

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

$$TT\left(P,t\right)=E\left(Y\left(P,t\right)-Y\left(P\right)|T=t\right)$$

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 who were retained in kindergarten. We can also define many more mean treatment parameters like the average effect of receiving treatment at $T=t$ versus receiving treatment at $T=t'$

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

or the effect of treatment at $T=t$ versus treatment at $T=t'$ for people who are actually treated at time $T = t''$

$$TT\left(P,t,t',t''\right) = E\left(Y\left(P,t\right) - Y\left(P,t'\right)|T = t''\right),$$

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 $\Phi\left(P,T\right)$ and/or the unobserved component of the outcome $\epsilon_i\left(P,T\right)$ depend on T or not, the effect of treatment will be dynamic. In the same manner, depending on whether $\epsilon_i\left(P,T\right)$ varies across individuals, the effect will be heterogeneous in the population. While under certain assumptions that limit the heterogeneity of treatment effects some of these parameters may equal one another, we consider the more general case where the treatment effect is allowed to vary over time and by unobserved individual characteristics.

3 Identification

The primary challenge to identifying treatment effects in the static framework lies in the fact 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 a higher ability student leading to a different effect of grade repetition. The challenge is similar in our dynamic setting, with the added challenge that selection is dynamic and that treatment effects vary both by unobservable type of individual and over time. To illustrate the challenges associated with identifying treatment effects in a dynamic context, we begin by considering what treatment effects can be recovered with experimental data, where the selection problem is eliminated by randomly assigning individuals to treatment. We then turn to the case of observational data. We consider how to extend several approaches that are used in the static framework to account for selection, namely control function and instrumental variable methods, and some of the shortcomings of these methods. Finally, we discuss an alternative approach of modeling the selection process by imposing a factor structure.

3.1 Experimental Data

Consider designing an experiment to recover some of the different population average parameters described above.[10] Consider first the case in which we are interested in estimates of ATE-type parameters. In this case, we can simply randomize people at the beginning of the first period into receiving treatment at each different possible treatment time (or not at all). While straightforward to recover, for the case of grade retention and arguably in many other applications as well, ATE-type parameters may not be particularly interesting from a policy perspective. For instance, in practice students who are retained are likely to have a higher potential benefit than the average

[10] While we focus on the binary treatment case, all of the points we make apply *mutatis mutandis* for the case in which, at any point in time, multiple treatments are possible.

student. Focusing on the average treatment effect would then bias us away from finding a positive effect of retention, even though it may be beneficial for lower-type students.

Treatment parameters that condition on the selection process (treatment on the treated and treatment on the untreated type parameters) are less straightforward to recover through random assignment to treatment and control groups. To illustrate we consider an example where treatment can be taken either of the first 2 periods $T = \{1, 2\}$, i.e., 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 3: $Y(3, T)$, e.g., third grade test scores. Let $R_i = 0$ if an individual is randomized into not receiving treatment and $R_i = T$ if the individual is randomized into receiving treatment at time T. Table 1 summarizes the experimental design for this case.

In period 1 individuals are selected into treatment or go on to the next period without treatment according to whatever selection process operates regularly (i.e. according to whether $V(1) > 0$ or $V(1) < 0$). Then, we take the individuals who would under normal circumstances receive treatment $T = 1$ (i.e. $V(1) > 0$) and randomize them into receiving treatment at $P = 1$, at $P = 2$ or not receiving treatment. In terms of our example, we observe children who would be retained in kindergarten and who would not according to some decision rule. Then, we take the students who would have been retained in kindergarten and randomly assign them to being retained in kindergarten, retained in first grade, or not being retained. From this randomization we are able to form all of the counterfactual outcomes conditional on $T = 1$ $(V(1) > 0)$.

We then go on to next period and we let those individuals who were not selected into treatment at 1 $(V(1) < 0)$ be selected into either $T = 2$ $(V(2) > 0)$ or into no treatment $(V(2) < 0)$. We then randomize them into either receiving treatment or not. We cannot randomize people into getting treatment $T = 1$ (elements in bold in Table 1) since that can only be done in period 1, but at period 1 we did not know whether they would be selected into $T = 2$ or $T = 0$. In other words, for a student who is selected to be retained in first grade, we cannot go back in time and randomly assign her to being retained in kindergarten and similarly for a student who is not selected into being retained in first grade. These mean outcomes are information that cannot be recovered because of the sequential nature (i.e. the dynamics) of the selection process. This means that we cannot address whether a student who is retained in first grade would have performed better if retained in kindergarten instead.

What this simple example shows is that, even in the scenario in which we can design an experiment to estimate mean treatment parameters, potentially policy-relevant information is lost. Some counterfactuals are lost because of the sequential nature of the selection process. Furthermore, the data requirements are much larger than in the static case, due to sample size requirements (i.e. the many randomizations across subgroups over time), and they get worse as the number of periods and/or the number of treatments is increased.

3.2 Observational Data

The randomized control trial provides a helpful starting point for considering how methods applied to account for selection in observational data in the static setting can be extended to a dynamic setting. We focus on the case where returns are heterogeneous both because this case is arguably empirically more relevant and because applying instrumental variables methods under homogeneity is a straightforward GMM problem.[11] In order to help fix ideas, we continue with our simple 3 period example and now ask whether methods designed to deal with the confounding effects of selection in observational data, namely control function and instrumental variables, will work in the dynamic heterogeneous case.

The potential outcomes in period 3 are given by

$$Y_i(3,T) = \Phi(3,T) + \epsilon_i(3,T) \text{ for } T = 0,1,2$$

and the observed outcome can be written as

$$\begin{aligned} Y_i(3) &= \Phi(3,0) + D_i(1)[\Phi(3,1) - \Phi(3,0)] + D_i(2)[\Phi(3,2) - \Phi(3,0)] \\ &\quad + \epsilon_i(3,0) + D_i(1)[\epsilon_i(3,1) - \epsilon_i(3,0)] + D_i(2)[\epsilon_i(3,2) - \epsilon_i(3,0)]. \end{aligned} \tag{3}$$

The (observed) outcome equation is a standard 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 et al. (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, which is likely in our case. That is, $D_i(T)$ is likely to be correlated with $\epsilon_i(3,T) - \epsilon_i(3,T')$ for $T \neq T'$. In the retention example, essential heterogeneity exists if students who would experience higher gains from retention are more likely to be retained.

3.2.1 Instrumental Variables

Consider first whether instrumental variables techniques can be applied to recover parameters of interest. First, recall that standard instrumental variable methods are invalid under essential heterogeneity even if we can find a Z that is statistically independent. Take, for example, the second unobservable term in equation (3). In this case we have

$$E(D_i(1)[\epsilon_i(3,1) - \epsilon_i(3,0)]|Z) = E(\epsilon_i(3,1) - \epsilon_i(3,0)|Z, D_i(1) = 1)\Pr(D_i(1) = 1|Z)$$

in the unobservables. Even though we assume $E(\epsilon_i(3,T)|Z) = 0$ for all T, $E(\epsilon_i(3,1) - \epsilon_i(3,0)|D_i(1) = 1, Z)$ will usually not equal zero since now we are also conditioning on $D_i(1) = 1$ and the decision

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

to get treatment is correlated with the unobservable gains associated with the treatment.

Instead of estimating ATE or TT like parameters one can address the problem of essential heterogenity within the instrumental variables framework by using local methods like the Local Average Treatment Effect ($LATE$) of Imbens & Angrist (1994)) and regression discontinuity designs (Hahn et al. (2001)). These methods deal with the problem of essential heterogeneity by recovering a "local" treatment parameter defined by some exogenous variation (e.g. an instrument that takes two values or a law that determines an exogenous cutoff) such that people affected by this variation are assigned into treatment independently of their potential outcomes. For example, in some states a child has to repeat a school grade if his test scores are below some cutoff. This kind of variation has been used in a regression discontinuity design (see Jacob & Lefgren (2004) and Nagaoka & Roderick (2005)) in which children just above and just below the cutoff are compared to estimate the effect of grade retention for children around the cutoff, the local treatment effect for this subgroup of students.

By definition, these methods will work in the presence of dynamic treatment effects, but one has to be careful both with the interpretation of the parameter they recover and with their implementation. The fact that we cannot recover the missing counterfactuals will have implications for what these local methods can actually recover. Consider our simple 3 period case and take the local average treatment effect as an example. Assume first that treatment is static by imposing that it can only be received at time $T = 1$ but not at $T = 2$. Assume also that we have an instrument Z that affects the choice of whether to receive treatment at time $T = 1$ but does not affect the outcomes. Furthermore, assume that Z can take two values, $z_2 > z_1$ such that $E\left(D_i\left(1\right)|Z = z_2\right) > E\left(D_i\left(1\right)|Z = z_1\right)$ for all i (i.e. the monotonicity condition of Imbens and Angrist). That is, individuals can only be induced into (but not out of) treatment when the instrument moves from z_1 to z_2 . Let $D_i\left(1, z_2\right)$ be the indicator of whether an individual gets treatment at period 1 when $Z = z_2$ and define $D_i\left(1, z_1\right)$ accordingly. In this binary treatment case the LATE parameter is given by

$$\begin{aligned} LATE\left(z_1, z_2\right) &= \frac{E\left(Y_i\left(3\right)|Z = z_2\right) - E\left(Y_i\left(3\right)|Z = z_1\right)}{E\left(D_i\left(1\right)|Z = z_2\right) - E\left(D_i\left(1\right)|Z = z_1\right)} \\ &= \frac{E\left(Y_i\left(3\right)|D_i\left(1, z_2\right) = 1\right) - E\left(Y_i\left(3\right)|D_i\left(1, z_1\right) = 0\right)}{E\left(D_i\left(1\right)|Z = z_2\right) - E\left(D_i\left(1\right)|Z = z_1\right)} \end{aligned}$$

so it measures the effect of treatment for those individuals induced into treatment by the change in the instrument.

Now suppose that individuals who are not affected by the instrument today can receive treatment in the next period, i.e. at time $T = 2$. The event $D_i\left(1, z_1\right) = 0$ will now include two types of individuals not induced into treatment at time 1: those who do not receive treatment at 2 still and those who receive treatment at time 2. Furthermore, while in the static case noncompliers, i.e., inframarginal individuals for whom $D_i\left(1, z_2\right) = D_i\left(1, z_1\right) = 0$, drop from the LATE calculation, in the dynamic case it may be the case that $D_i\left(2, z_2\right) \neq D_i\left(2, z_1\right)$. LATE will now be a weighted (by the probabilities of each of these events) average of these different kinds of individuals and harder to interpret.[12]

By imposing strong restrictions on the selection process (mainly that $U_i(T) = U_i$ for all T), an alternative is the local instrumental variables approach to ordered choice models of Heckman & Vytlacil (2007) that recovers pairwise Marginal Treatment Effects (MTE). Using the MTE some of the missing TT-type parameters can be recovered. Alternatively if one has access to very special kind of data one can be relatively agnostic about the selection process. Nekipelov (2008), for example, uses a multivalued instrument that satisfies a different kind of monotonicity: as the value of the instrument increases people either do not change treatment at all or they change treatment monotonically. This avoids the problem with th estandard LATE approach described above at the cost of requiring a very particular type of instrument and decision process.

3.2.2 Control Function

An alternative to instrumental variables methods is to use the control function approach which models the selection process explicitly,[13] extending it to account for dynamics. Let $P_{i,1}$ denote the probability of getting treated at $T = 1$. The event $T = 1$ can be written as

$$\begin{aligned} U_i(1) > -\lambda(1) \quad &\iff \quad F_{U(1)}(U_i(1)) > F_{U(1)}(-\lambda(1)) \\ &\iff \quad F_{U(1)}(U_i(1)) > 1 - P_{i,1}, \end{aligned}$$

i.e. as a function of $P_{i,1}$. Next, form the observed conditional mean of outcome 1 when $T = 1$ and rewrite

$$\begin{aligned} E(Y_i(3,1)|T=1) &= \Phi(3,1) + E(\epsilon_i(3,1)|T=1) \\ &= \Phi(3,1) + E(\epsilon_i(3,1)|U_i(1) > -\lambda(1)) \\ &= \Phi(3,1) + K_1(P_{i,1}). \end{aligned}$$

The term $K_1(P_{i,1})$ is known as a control function and it can be identified nonparametrically under various conditions. The simplest condition is when one has exclusion restrictions, i.e., instrumental variables that affect the probability of getting treatment but not the outcome of interest directly. As shown in Heckman & Navarro (2007) other nonparametric restrictions are possible. Once $K_1(P_{i,1})$ is recovered one can apply the law of iterated expectations to get

$$E(\epsilon_i(3,1)) = K_1(P_{i,1})P_{i,1} + E(\epsilon_i(3,1)|T \neq 1)(1 - P_{i,1}) = 0.$$

The only unknown term in this expression is $E(\epsilon_i(3,1)|T \neq 1)$ so we can solve for it. However, as with the case of experimental data neither $E(\epsilon_i(3,1)|T=0)$ nor $E(\epsilon_i(3,1)|T=2)$ can be recovered. Because $T = 2$ is the terminal treatment in this example, all the remaining counterfactuals that can be recovered with experimental data can also be recovered with the control function by

[12] See Angrist & Imbens (1995) for a similar result in a model with multiple treatments.

[13] See Heckman & Robb (1985) and Navarro (2008).

using a similar reasoning (i.e. by forming control functions for $T = 0$ and $T = 2$ which will be functions of $P_{i,1}$ and $P_{i,2}$ and proceeding sequentially using the law of iterated expectations).

Using a control function approach one can take advantage of the availability of instruments, allow for essential heterogeneity, and recover the same treatment parameters of interest as in a randomized trial. Notice that modelling the selection process does not overcome the problem of the missing counterfactuals. In order to recover these additional counterfactuals further assumptions on the joint distribution of the unobserved components, like the factor structure we present below, are needed.

3.2.3 Factor Structure

We now propose a less traditional solution to selection by imposing a factor structure that permits us to recover the joint distribution of the unobservables. This not only solves the missing counterfactuals problem, but also incorporates some other nice features that are discussed further below. In particular, we impose the following assumption

A-2 *(Factor structure)* $\epsilon_i(P,T) = \theta_i \alpha(P,T) + \varepsilon_i(P)$ and $U_i(T) = \theta_i \rho(T) + \upsilon_i(T)$ where θ_i is a vector of mutually independent "factors" and we assume that $\varepsilon_i(P) \perp\!\!\!\perp \varepsilon_i(P')$ for all $P \neq P'$ and $\upsilon_i(T) \perp\!\!\!\perp \upsilon_i(T')$ for all $T \neq T'$ where $\perp\!\!\!\perp$ denotes statistical independence[14].

We impose **A-2** for convenience even though it is stronger than required.[15] The factor structure assumption is a convenient dimension reduction technique. It allows us to transform the enormous problem of identifying and estimating the joint distribution of all the unobservables (U_i, ϵ_i) into a simpler problem: that of recovering the factor "loadings" $\alpha(P,T)$ and $\rho(T)$ and the marginal distributions of the elements of θ_i and of $\varepsilon_i(P,T), \upsilon_i(T)\ \forall P,T$.

Not only is the factor structure convenient, it also aids in interpretation of results since we can now talk about a low dimensional set of common "causes."[16] Furthermore, it is intuitively appealing in the way it interprets the role of heterogeneity; namely, that the same set of unobservables (the vector θ_i) that determines the effect of treatment also determines the selection into treatment. In our grade retention example, if θ_i is unobserved ability (or abilities if θ is a vector), essential heterogeneity arises because unobserved ability affects both the treatment effect (i.e., gain in test scores across two consecutive 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.

To understand how the factor structure addresses the identification problem associated with unobserved heterogeneity, consider our three period example. If **A-2** holds, the observed outcome

[14] If **A-1** holds, $\alpha(P,T) = \alpha(P,0) = \alpha(P)$ for $T \geq P$.

[15] Following the analysis of measurement error models in Schennach (2004) we can relax the strong statistical independence assumptions and replace them with a combination of general dependence and weaker mean independence assumptions.

[16] See Jöreskog & Goldberger (1975) for a discussion and Carneiro et al. (2003) and Cunha et al. (2005) for recent developments.

will be

$$\begin{aligned} Y(3) &= \Phi(3,0) + D_i(1)\left[\Phi(3,1) - \Phi(3,0)\right] + D_i(2)\left[\Phi(3,2) - \Phi(3,0)\right] + \varepsilon(3) \\ &\quad + \theta_i \alpha(3,0) + D_i(1)\theta_i\left[\alpha(3,1) - \alpha(3,0)\right] + D_i(2)\theta_i\left[\alpha(3,2) - \alpha(3,0)\right], \end{aligned}$$

and the choice process will be determined by

$$V_i(T) = \lambda(T) + \theta_i \rho(T) + \upsilon_i(T).$$

In this case essential heterogeneity is present when $\alpha(3,T) \neq \alpha(3,0)$ since now the unobserved gains in the test score will be correlated with the choice indicator because the same θ_i determines both.

If we could recover (or condition on) the unobserved θ_i, the selection process $V_i(T)$ and the outcome of interest $Y(P,T)$ would be independent and we could then 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 quasiexperimental methods. There are many normalizations under which the distribution of θ_i can be recovered (see Cunha et al. (2006) and Robin & Bonhomme (2008) for examples).

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 unidimensional. Suppose the outcome in period 1 is free of selection,[17] so

$$Y(1) = \Phi(1) + \theta_{i,1}\alpha_1(1) + \varepsilon(1).$$

It is straightforward to show that the joint distribution of $\epsilon_i(1) = \theta_{i,1}\alpha_1(1) + \varepsilon(1)$ and $U_i(1) = \theta_{i,1}\rho_1(1) + \upsilon_i(1)$ is nonparametrically identified (e.g. Heckman & Smith (1998)). >From it, normalizing $\rho_1(1) = 1$,[18] 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 by Kotlarski (1967) that the distribution of $\theta_{i,1}$ (and of $\varepsilon(1)$ and $\upsilon(1)$) is nonparametrically identified. Intuitively, we can, from $E\left(\epsilon_i^k(1)\,U_i(1)\right) = \alpha_1^k(1)\,E\left(\theta_{i,1}^{k+1}\right)$ for $k > 0$, recover all the moments of $\theta_{i,1}$. Since we can recover all moments of the random variable $\theta_{i,1}$ we can, for all practical purposes, recover its distribution. Formally, one wants to characterize a distribution using its characteristic function and not moments since some distributions are not characterized by their moments (see Casella & Berger (2002) for conditions).

[17] 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 & Smith (1998).

[18] 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.

This is precisely what the Kotlarski argument does.

Next consider the (selection corrected) second period equations

$$Y\left(2,T\right)=\Phi\left(2,T\right)+\theta_{i,1}\alpha_{1}\left(2,T\right)+\theta_{i,2}\alpha_{2}\left(2,T\right)+\varepsilon\left(2,T\right)\text{ for }T=0,1$$

$$V_{i}\left(2\right)=\lambda\left(2\right)+\theta_{i,1}\rho_{1}\left(2\right)+\theta_{i,2}\rho_{2}\left(2\right)+\upsilon_{i}\left(2\right),$$

where we now allow for a new element of θ_i to enter. $\theta_{i,2}$ can be interpreted as a correlated shock, i.e. an unobserved shock that affects outcomes and selection equations from period 2 on although 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. Identification is straightforward. By taking cross moments over time (i.e. $Y\left(1\right)$ with the selection corrected $Y\left(2,T\right)$) one can identify the elements associated with $\theta_{i,1}$ in period 2 equations. Then, by taking cross moments within period 2 equations, one can identify the elements associated with the correlated shock $\theta_{i,2}$ as well as the nonparametric distributions of the unobservables.

We can extend this analysis to the case in which unobserved ability (θ_i) is multidimensional beyond the correlated shocks. For example, in our empirical application we consider a normalization of θ_i that is particularly relevant to retention decisions when students may be retained both for cognitive and behavioral abilities, but is likely to be applicable to other settings where the unobservable component is multidimensional. Thus, we propose that true ability consists of three independent components: 1) A trait associated purely with cognitive functions C, a purely behavioral trait B and general ability A that can be used for both cognitive and behavioral functions. That is, we assume that $\theta_i=(A_i,B_i,C_i)$.[19] Associated with ability is a set of tests or markers that measure these components of ability imperfectly, but are free of selection. In our empirical example, these correspond to the initial tests applied to students in kindergarten before any grade repetition takes place. This requirement is not crucial (provided we can correct for selection) but we keep it because a) it is common to many situations and b) it makes the exposition of the identification argument much simpler.[20]

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.[21] We write the j^{th} demeaned cognitive test as

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

and the j^{th} demeaned behavioral test as

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

[19] While we write the discussion in terms of test scores and abilities, the normalization we present applies generically to other situations. Other normalizations are possible.

[20] Heckman & Navarro (2007) show it is not required.

[21] And we assume that these covariates are independent of θ_i. It will be clear below that full independence is stronger than required, conditional moment independence for certain moments is enough.

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 and general ability A and some (like measures of class disruptive behaviors or habits) are noisy measures of the behavioral trait B and general ability A. 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. The cognitive ability component C and the behavioral component B measure the part of ability that is used exclusively for the corresponding function.

Semiparametric identification follows from an argument similar to the one used for the one factor model but now we take moments across cognitive and behavioral equations and then within cognitive test and within behavioral test to recover the α, π and ϕ parameters as well as the nonparametric distributions of A, 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

$$\begin{array}{c} E\left(\left(\zeta_j\right)^n \beta_k\right) = \alpha_{\zeta,j}^n \alpha_{\beta,k} E\left(A^{1+n}\right) \\ E\left(\zeta_j \left(\beta_k\right)^h\right) = \alpha_{\zeta,j} \alpha_{\beta,k}^h E\left(A^{1+h}\right) \end{array} . \tag{4}$$

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

$$\frac{\zeta_j}{\alpha_{\zeta,j}} = \left[C \frac{\pi_{\zeta,j}}{\alpha_{\zeta,j}} + \frac{\varepsilon_{\zeta,j}}{\alpha_{\zeta,j}}\right] + A,$$

$$\frac{\beta_k}{\alpha_{\beta,k}} = \left[B \frac{\phi_{\beta,k}}{\alpha_{\beta,k}} + \frac{\varepsilon_{\beta,k}}{\alpha_{\beta,k}}\right] + A.$$

Then, it follows from a Theorem by Kotlarski (1967) that 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. Intuitively, given the now known $\alpha_{\zeta,j}$ and $\alpha_{\beta,k}$, we can identify all of the moments of general ability A from equation (4). Since we can recover all moments of the random variable A we can, for all practical purposes, recover its distribution. Formally, one wants to characterize a distribution using its characteristic function and not moments

[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.

since some distributions are not characterized by their moments (see Casella & Berger (2002) for conditions). This is exactly what the Kotlarski argument does.

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 forming

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

we can recover $\pi_{\zeta,k}$ for all $k=2,...,N_c$. By iteratively applying the Kotlarski argument, we can nonparametrically recover the distributions of C and $\varepsilon_{\zeta,j}$ for all $j=1,...,N_c$. 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 all $j=1,...,N_b$.

Once we have recovered the distribution of θ_i, we can proceed to the next period. Now some children will be treated (i.e. will repeat first grade) 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[23] like the control function approach[24] We can then repeat the arguments above and recover the loadings and the distribution of the uniquenesses. However, since we now know the distribution of abilities in advance, we can let all three types of ability enter all equations (whether behavioral or cognitive) without having to normalize some loadings to zero. That is, the normalization imposing that B only enters β equations and C only enters ζ equations, need only apply on 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. However, 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. Formally, consider a modified version of the model of equations (**??**) and (**??**) in a multiperiod setting. In period 1 the model is given by:

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

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

[23] Ideally with access to an exclusion restriction in order to attain nonparametric identification as in Heckman (1990) and Heckman & Smith (1998).

[24] See Heckman & Robb (1985) and Navarro (2008).

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 τ are writen as

$$\zeta_{i,k,\tau,t} = A_i \alpha_{\zeta,k,\tau,t} + B_i \phi_{\zeta,k,\tau,t} + C_i \pi_{\zeta,k,\tau,t} + \sum_{h=2}^{t} \eta_i^{(h)} \delta_{\zeta,k,\tau,t}^{(h)} + \varepsilon_{\zeta,k,t}. \tag{5}$$

First, notice that we now allow for behavioral ability to potentially determine cognitive tests after period 1. Second, we also add a new unobservable $\eta_i^{(h)}$ every period. Since this new unobservable is individual specific and we allow it to affect all outcomes (and retention decisions) from period h on, it can be interpreted as a permanent shock that hits the model in period h (hence the superscript). While the shock itself is permanent we allow for its effects to change both over time and across retention status for all equations in the model.

Now consider identification of equation (5) in period 2 for retention status τ. We can form cross second moments between period 2 and period 1 cognitive tests:

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

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

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

and recover the behavioral ability loadings for all $k = 1, ..., N_c$ and all retention statuses.

In order to identify the terms related to the new unobservable (i.e. the period 2 permanent shock $\eta^{(2)}$ and its loadings $\delta_{\zeta,k,\tau,2}^{(2)}$) a normalization on the scale of the unobservable is required so we impose that $\delta_{\zeta,1,0,2}^{(2)} = 1$ for $\tau = 0$. With the normalization in place we can form cross moments between period 2 equations for the $\tau = 0$ retention status and form

$$\frac{E\left(\zeta_{k,0,2}, \zeta_{k',0,2}\right) - \alpha_{\zeta,k,0,2}\alpha_{\zeta,k',0,2} E\left(A^2\right) + \pi_{\zeta,k,0,2}\pi_{\zeta,k',0,2} E\left(C^2\right)}{E\left(\zeta_{1,0,2}, \zeta_{k',0,2}\right) - \alpha_{\zeta,1,0,2}\alpha_{\zeta,k',0,2} E\left(A^2\right) + \pi_{\zeta,1,0,2}\pi_{\zeta,k',0,2} E\left(C^2\right)} = \delta_{\zeta,k,0,2}^{(2)}$$

to identify the loadings on the permanent shock for all cognitive scores $k = 1, ..., N_c$ for retention status $\tau = 0$.[25]We can then apply Kotlarsky to any pair of equations k, k' for $\tau = 0$ and identify the nonparametric distributions of $\eta^{(2)}$ and $\varepsilon_{\zeta,k,2}, \varepsilon_{\zeta,k',2}$. To identify the loadings for retention statuses

[25] Notice that we cannot form cross moments for equations with different retention indices τ since we can only observe a kid in the retention status he actually receives.

$\tau > 0$ we can form

$$\frac{E\left(\zeta_{k,\tau,2},\zeta^2_{k',\tau,2}\right)-\alpha_{\zeta,k,\tau,2}\alpha^2_{\zeta,k',\tau,2}E\left(A^3\right)+\pi_{\zeta,k,\tau,2}\pi_{\zeta,k',\tau,2}E\left(C^3\right)}{E\left(\zeta_{k,\tau,2},\zeta_{k',\tau,2}\right)-\alpha_{\zeta,k,\tau,2}\alpha_{\zeta,k',0,2}E\left(A^2\right)+\pi_{\zeta,k,\tau,2}\pi_{\zeta,k',0,2}E\left(C^2\right)}\frac{E\left(\left(\eta^{(2)}\right)^2\right)}{E\left(\left(\eta^{(2)}\right)^3\right)}=\delta^{(2)}_{\zeta,k,\tau,2}.$$

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

4 The Effect of Retention on Test Scores

Grade retention is a common and controversial practice in U.S. schools. Understanding the effects of grade retention has become increasingly important with the rise in student accountability policies that often include grade retention as a potential consequence of not meeting a given achievement threshold. Most research on the effects of grade retention treats it as a single treatment (being retained versus not being retained) or attempts to correct for static, but not dynamic, selection. These studies generally find that retention at best has no effect and at worse has considerable negative effects.[26]

We apply the method described above to estimate how the effect of grade retention varies across different grades, as time since retention passes and by different types of students using data from the Early Childhood Longitudinal Survey (ECLS-K). The ECLS-K is a panel survey of students starting with the 1998-99 kindergarten cohort. The survey was applied again in the 1999-2000, 2001-02 and 2003-04 school years. 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. 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 and fourth grades).[27]That is, our dynamic treatment time indicator takes values $T = 0, 1, 2, 3$ where $T = 0$ means the child was not retained, $T = 1$ means he is retained in kindergarten, $T = 2$ that he is retained early and $T = 3$ that he is retained late.

Each year of the ECLS-K includes cognitive tests measuring students' reading and math skills as well as teacher ratings on students' behavioral and social skills–the Social Rating Scale (SRS). We focus primarily on effect of retention on the cognitive tests, the item response theory (IRT)

[26] See Holmes (1989) and Jimerson (2001) for comprehensive meta-analyses. In more recent studies, using a regression discontinuity design to study test-based promotion in Chicago public schools, both Jacob & Lefgren (2004) and Nagaoka & Roderick (2005) found that retention lead to small short term gains on test scores that disappeared over time.

[27] In principle we could form all and separate early and late into the four grades. This, however, can only be done for less than half of the sample.

scores, as our outcomes of interest, though the effect of retention on behavior is of interest as well.

A logical difficulty in evaluating the effect of grade retention is that it is impossible to hold both the grade and age fixed when retaining a 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. 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 survey also includes information on the schools' retention policies in all survey years. We use these variables as exclusions, under the assumption that they do not determine the child's test score directly (conditional on the other covariates including observable school characteristics) but they do affect the probability that a child repeats a grade.

The ECLS-K contains a very rich set of covariates that include characteristics of the children, the family, the class and the school that we use as controls. Table 2 shows descriptive statistics for the covariates we include in all our equations as well as the the retention policy variables for the first year of the survey for the overall sample and broken out by whether the student is ever retained. There are 576 students in our base year sample who are retained either in kindergarten, early or late. 64% of students who are retained are male, compared to about 49% in the overall sample. Students who are retained are also more likely to be racial minorities, have more siblings, lower SES, and no father in the home. Class and teacher characteristics are similar across the two samples, with the exception that retained students come from classrooms with larger percentages of minorities than non-retained students. Retained students also appear to come from schools with higher percentages of minorities and where security is more of an issue than non-retained students.

Tables 3 and 4 compare student achievement in reading and math, respectively across time and retention statuses. Not surprisingly, students who are retained have lower test scores in both reading and math than students who are not retained. Students who are retained in kindergarten have slightly higher test scores in kindergarten than students who are retained early, but lower test scores than students who are retained late. By the next year of the survey, when students would be in first grade in the absence of retention, students who are retained in kindergarten have lower test scores than any of the other subgroups. Interestingly, by the time students would be in third grade in the absence of retention, students who were retained early (i.e., in first or second grade) have the lowest test scores. Furthermore, the students who are retained late (i.e., in third or fourth grade) have lower test scores in the third grade year than students who were retained in kindergarten (and are now in second grade). While not conclusive, these patterns in test scores suggest some trends that may be consistent with both dynamic selection and time-varying treatment effects.

Tables 5 to 7 show similar patterns in the students' behavioral measure by year and retention status. For instance, students who are retained early have the lowest behavioral scores of all the retention status in kindergarten and in 1999/00 when students would be in first grade in the absence of retention. Students who are retained late have the lowest behavioral score in 2001/02 when students would be in third grade in the absence of retention. We only use the SRS scores to control for behavioral ability in the selection equation which is only estimated prior to the last wave, so we do not include measures for 2003/04.

4.1 Baseline Estimates

To demonstrate the potential importance of our method, we present suggestive evidence on both the presence of dynamic selection and of dynamic treatment effects using the ECLS-K. To test for dynamic selection, we regress the kindergarten cognitive tests on indicators of whether the child is retained in the future, controlling for the covariates related to the child, his family, school and class as described in Table 2 above. Tables 8 and 9 present results for math and reading respectively. Column 1 of both tables show that children who will be retained have lower kindergarten test scores than those who will not be retained. Furthermore, we reject the hypothesis that the effects of being retained at different grades are the same, suggesting not only the presence of selection but also *dynamic* selection on cognitive test scores.

We then look for evidence of dynamic treatment effects, by regressing test scores in the last sample period (2003/04 school year) on retention in the different periods. As shown in Column 2 of Tables 8 and 9, being retained is associated with worse outcomes. The coefficients on the different retention statuses are also significantly different from each other. This could be because treatment effects vary over time, or because of selection of different types of students into retention at different grades.

One way to begin to control for a static component of selection is to include achievement in kindergarten, prior to any retention decisions taking place. Columns 3 and 4 present results controlling for adding in 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. Furthermore, we can reject the formal test of equality of the effects for different retention times.[28]

While this basic analysis provides suggestive evidence of both dynamic 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 treatment effects estimated.[29] Furthermore, this analysis does not capture heterogeneous effects of treatment by student type, which is a central contribution of our paper.

[28] As before, the same pattern holds for the other cognitive tests and for the behavioral measures.

[29] See Heckman & Navarro (2004).

4.2 Estimating a Multidimensional Model of Ability and Retention

Following our discussion of identification in Section 3.2.3, we impose the following normalizations. We normalize the general ability loading on the general knowledge test to 1, so A can be interpreted as a trait that is associated positively with higher scores in the general knowledge test. The loading on cognitive ability is normalized to 1 on the math test, so C is associated with higher math scores. Finally, we normalize the behavioral loading on the self-control marker to 1. If we 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

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

and

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

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

$$\begin{aligned} \zeta_{i,j,t} \quad = \quad & \textstyle\sum_{\tau=1}^{3} D\left(\tau\right)\Big[\Phi_{\tau,t} + A_i\left[\alpha_{\zeta,j,\tau,t} - \alpha_{\zeta,j,t}\right] + B_i\left[\phi_{\zeta,j,\tau,t} - \phi_{\zeta,j,\tau,t}\right] + C_i\left[\pi_{\zeta,j,\tau,t} - \pi_{\zeta,j,t}\right] + \sum_{h=2}^{t}\eta_i^{(h)}\Big[\delta_{\zeta,j,\tau,t}^{(h)} - \\ & + X_{i,t}\gamma_{\zeta,j,t} + A_i\alpha_{\zeta,j,t} + B_i\phi_{\zeta,j,t} + C_i\pi_{\zeta,j,t} + \textstyle\sum_{h=2}^{t}\eta_i^{(h)}\delta_{\zeta,j,\tau,t}^{(h)} + \varepsilon_{\zeta,j,t} \end{aligned} \tag{8}$$

Importantly, note that this specification corresponds to the general case discussed above, in that the treatment varies over time as does the marginal effect of observable characteristics and unobservable "abilities."

The decision to have a child repeat a grade 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 & Navarro (2007), this model is in fact nonparametrically identified and it follows from the same arguments used in section 3.2.3.

The actual form of the model for retention we use is the following. We write the latent index V as[30]

$$V_{i,t} = \lambda_{0,t} + X_{i,t}\lambda_x + Z_{i,t}\lambda_z + A_i\rho_{A,t} + B_i\rho_{B,t} + C_i\rho_{C,t} + \sum_{h=2}^{t}\eta_i^{(h)}\psi_t^{(h)} + \upsilon_{i,t} \text{ for } t = \underline{T}, ..., \bar{T}.$$

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

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

Notice that, consistent with our data, we allow for exclusions in the index. That is we allow for some variables (Z) to be included in the retention equations but not in the outcomes. In the data

[30] 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 471, and the number of parameters would increase considerably, we stick with the linear form.

this correspond to 7 binary measures of 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, these policies vary considerably across schools. In the 1998-1999 school year 75 percent of schools permit retention by parental request, and only 45 percent permit retention without parental permission.

As shown in section 3.2.3, the distributions of the unobservables $(A, B, C, \{\eta^{(h)}\}_{h=2}^{T}, \varepsilon, \upsilon)$ in the model are nonparametrically identified. For estimation purposes, however, we specify all of the distributions and allow them to follow mixtures of normals with either two or three components. Second, while our identification arguments are presented in a sequential fashion and lead naturally to a multistep estimation procedure we estimate all of the parameters in the model jointly by maximum likelihood in one single step.

4.3 Dynamic effects of retention

4.3.1 Model Fit

In Tables 10 and 11 we present evidence of the fit of the model. Table 10 shows that the model fits the means and variances of all the test measures very well, and we cannot reject that these are the values predicted by the model. Table 11 shows that the same is true for the probabilities of retention in the data. We cannot reject the hypothesis of equality of predicted and actual probabilities.

4.3.2 Unobservable Abilities

Figures 1 through 3 present evidence for selection on the components of ability. Figure 1 describes the distribution of general ability by retention status; Figure 2 corresponds to behavioral ability and Figure 3 to general ability. Not surprisingly, students who are not retained have higher general and cognitive ability than those who are retained. Students who are retained in kindergarten have generally higher behavioral and general ability than those retained early. Students retained late have higher behavioral and general ability than those who are retained earlier. It is more difficult to compare cognitive ability among retained students. This suggests that while cognitive ability is higher for non-retained students, it is not playing as much of a role as general and behavioral ability in distinguishing when students are retained.

4.3.3 Average Treatment Parameters

Tables 14 and 15 describe treatment on the treated (and the untreated) as well as unconditional average treatment parameters for both reading and math test scores in the 2003-04 school year when students are age 11. The columns correspond to actual treatment statuses, whereas the rows compare potential treatment statuses. In other words, the first row describes the treatment effect of being retained in kindergarten versus not being retained. The average treatment effects in the last column show that students perform 4% higher in reading and 13% higher in math by 2003-04 (period 4) if retained in kindergarten versus not being retained. The average treatment effect for early retention is smaller for both reading (2%) and math (5%). The average treatment effect of late retention is comparable to kindergarten retention and slightly higher, 6% for reading and 14% for math.

However, turning to the treatment on the treated parameters, in the first column we see that these positive average treatment effects are being driven by the potential positive effect on students who are not actually retained. A student who is retained in kindergarten performs 4% lower in math and reading by 2003-04 than if not retained. Students who are retained early perform 7% lower in reading and 5% lower in math. Students who are retained late would have performed better in math if not retained, 3%, but do marginally better in reading.

While we have focused the discussion on the return in terms of longer term outcomes, i.e., test scores in 2003-04, short term outcomes may also be of interest. We also consider the effect of retention on test scores in earlier periods. Figure 4 compares the treatment effect at the different time periods for reading and math. We focus on the effect of being retained in kindergarten versus not being retained. The left hand side figure shows that the average treatment effect of kindergarten retention is initially strongly negative for reading but becomes positive as time sense retention passes. The gains for kindergarten retention for students who are actually retained in kindergarten (the treatment on the treated) in the right hand panel show a similar pattern. However, this time both the treatment effect of retention in math and reading are strongly negative, about 19% lower for math and 26% for reading, and approach 0 by 2 years later and level off. As shown in the previous tables, the gains do not become positive. Interestingly, this appears to be in contrast to evidence in the literature which suggests that any gain in retention may actually be short-lived.[31]

Figure 5 shows the time trend for early retention. In this case, the average treatment effect for reading stays about 0 over time. The average treatment effect for math is initially negative, -10%, but becomes weakly positive 2 years later. The treatment on the treated, i.e., on those who are actually retained early, in the right hand panel shows a different pattern. The initial effect on reading achievement is negative and becomes less negative 2 years later. For math, the initial treatment effect is approximately 0, but becomes more negative over time. Because we only have one post-retention test score for students who are retained late, we cannot investigate trends in this case.

While comparing treatments across treatment statuses suggests that treatment may vary by student ability, Figures 6 to 8 provide further evidence about how the treatment effect varies by different measures of students abilities. Figure 6 describes how the treatment effect of being retained varies across the percentiles of the general, behavioral and cognitive ability distributions for reading and math by 2003/04. For reading and math, the benefits of retention increase with general and cognitive ability for across retention times, particularly for math with cognitive ability. For behavioral ability, late and early retention exhibit the opposite pattern, with lower ability students demonstrating larger improvements due to retention than higher ability students.

[31] See Frederick & Hauser (2006) for a nice summary of the literature.

Figures 7 and **??** present findings for earlier periods.

These apparently counterintuitive finding of the higher ability students benefiting from retention more than lower ability students (particularly in the long run) may be explained by first observing that the test scores reported in the ECLS-K are not actually those used to determine retention decisions. High ability students who are retained may benefit more than low ability students from that additional year if they benefit more from any additional attention given to them by their teachers or parents. This finding is further supported by research by Bedard & Dhuey (2006) and others suggesting that the age relative to other children in the classroom matters for performance. The disparity in findings across high and low ability students may simply follow because a lower ability student who is retained may be discouraged if he finds out that now younger students are outperforming him.

Importantly, this finding may help to reconcile some of the mixed evidence on the effects of grade retention in the literature. Effectively, 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 these marginal students may find a positive effect of retention even if lower-ability students are being hurt by the design.

5 Conclusion

In this paper we develop and apply a framework for the analysis of dynamic treatment effects. Our analysis of grade retention shows the usefulness of extending the standard static framework to estimate dynamic treatment effects. First, preliminary results show evidence of dynamic selection, which is not accounted for in previous studies in the literature. Second, preliminary estimates show that the effects of repeating a grade on tests scores at age 11 vary considerably depending on when the student is retained, i.e., dynamic treatment effects. Another benefit of our framework is that we can contrast the effects of retention not just across time but across students of different unobservable abilities. Interestingly, our preliminary results also suggest that generally higher ability students benefit more from retention than lower ability students.

The analysis in this paper could clearly be applied in other situations like treatments associated with health status indicators and/or costs of particular treatments. One could also analyze the effects of advertisement at different stages in the life of a product, the effect of attending a segregated school at different stages, and similar problems.

The framework we develop can be thought of as a midpoint between the standard reduced form static treatment literature and a fully specified structural dynamic discrete choice model. In many situations it is not clear how to specify the selection process and our analysis provides a reduced form alternative (with all the advantages and problems associated with it). Furthermore, since extending it to the case in which treatment is not an absorbing state is straightforward (by letting treatment occur not only the first time a threshold is crossed but also the second, third, etc) it can be applied in more complicated situations.

References

Abbring, J. H. & Van den Berg, G. J. (2003), 'The nonparametric identification of treatment effects in duration models', *Econometrica* **71**(5), 1491–1517.

Angrist, J. D. & Imbens, G. W. (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., Cooley, J. & Hussey, A. (2008), 'The economic return to an mba', *International Economic Review* **49**(3), 873 – 899.

Bedard, K. & Dhuey, E. (2006), 'The persistence of early childhood maturity: International evidence of long-run age effects', *The Quarterly Journal of Economics* **121**(4), 1437–1472.

Carneiro, P., Hansen, K. & Heckman, J. J. (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.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, 2nd edition edn, Duxbury Press.

Cunha, F., Heckman, J. J. & Navarro, S. (2005), 'Separating uncertainty from heterogeneity in life cycle earnings, the 2004 Hicks lecture', *Oxford Economic Papers* **57**(2), 191–261.

Cunha, F., Heckman, J. J. & Navarro, S. (2007), 'The identification and economic content of ordered choice models with stochastic cutoffs', *International Economic Review* **48**(4), 1273 – 1309.

Cunha, F., Heckman, J. J. & Schennach, S. M. (2006), Estimating the technology of cognitive and noncognitive skill formation. Unpublished manuscript, University of Chicago, Department of Economics. Presented at the Yale Conference on Macro and Labor Economics, May 5–7, 2006. Under revision, *Econometrica*.

Eberwein, C., Ham, J. C. & LaLonde, R. J. (1997), 'The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: Evidence from experimental data', *Review of Economic Studies* **64**(4), 655–682.

Frederick, C. B. & Hauser, R. M. (2006), Have we put an end to social promotion? Changes in grade retention rates among children aged 6 to 17 from 1972 to 2003.

Gill, R. D. & Robins, J. M. (2001), 'Causal inference for complex longitudinal data: The continuous case', *The Annals of Statistics* **29**(6), 1785–1811.

Hahn, J., Todd, P. E. & Van der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design', *Econometrica* **69**(1), 201–209.

Heckman, J. J. (1990), 'Varieties of selection bias', *American Economic Review* **80**(2), 313–318.

Heckman, J. J., Hotz, V. J. & Walker, J. R. (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. & Navarro, S. (2004), 'Using matching, instrumental variables, and control functions to estimate economic choice models', *Review of Economics and Statistics* **86**(1), 30–57.

Heckman, J. J. & Navarro, S. (2007), 'Dynamic discrete choice and dynamic treatment effects', *Journal of Econometrics* **136**(2), 341–396.

Heckman, J. J. & Robb, R. (1985), Alternative methods for evaluating the impact of interventions, *in* J. Heckman & B. Singer, eds, 'Longitudinal Analysis of Labor Market Data', Vol. 10, Cambridge University Press, New York, pp. 156–245.

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

Heckman, J. J., Urzua, S. & Vytlacil, E. J. (2006), 'Understanding instrumental variables in models with essential heterogeneity', *Review of Economics and Statistics* **88**(3), 389–432.

Heckman, J. J. & Vytlacil, E. J. (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* J. Heckman & E. Leamer, eds, 'Handbook of Econometrics, Volume 6', Elsevier, Amsterdam. Forthcoming.

Heckman, J. J. & Walker, J. R. (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* L. Shepard & M. Smith, eds, 'Flunking grades: Research and policies on retention', The Falmer Press, London, pp. 16–33.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475.

Jacob, B. & Lefgren, L. (2004), 'Remedial education and student achievement: A regression-discontinuity analysis', *Review of Economics and Statistics* **86**(1), 226–244.

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. & Goldberger, A. S. (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, Technical Report 2004, IZA Discussion Paper.

Matzkin, R. L. (1992), 'Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models', *Econometrica* **60**(2), 239–270.

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. & Roderick, M. (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* S. N. Durlauf & L. E. Blume, eds, 'The New Palgrave Dictionary of Economics.', second edn, Palgrave Macmillan Press, London.

Nekipelov, D. (2008), Endogenous multi-valued treatment effect model under monotonicity. Unpublished manuscript, Berkeley.

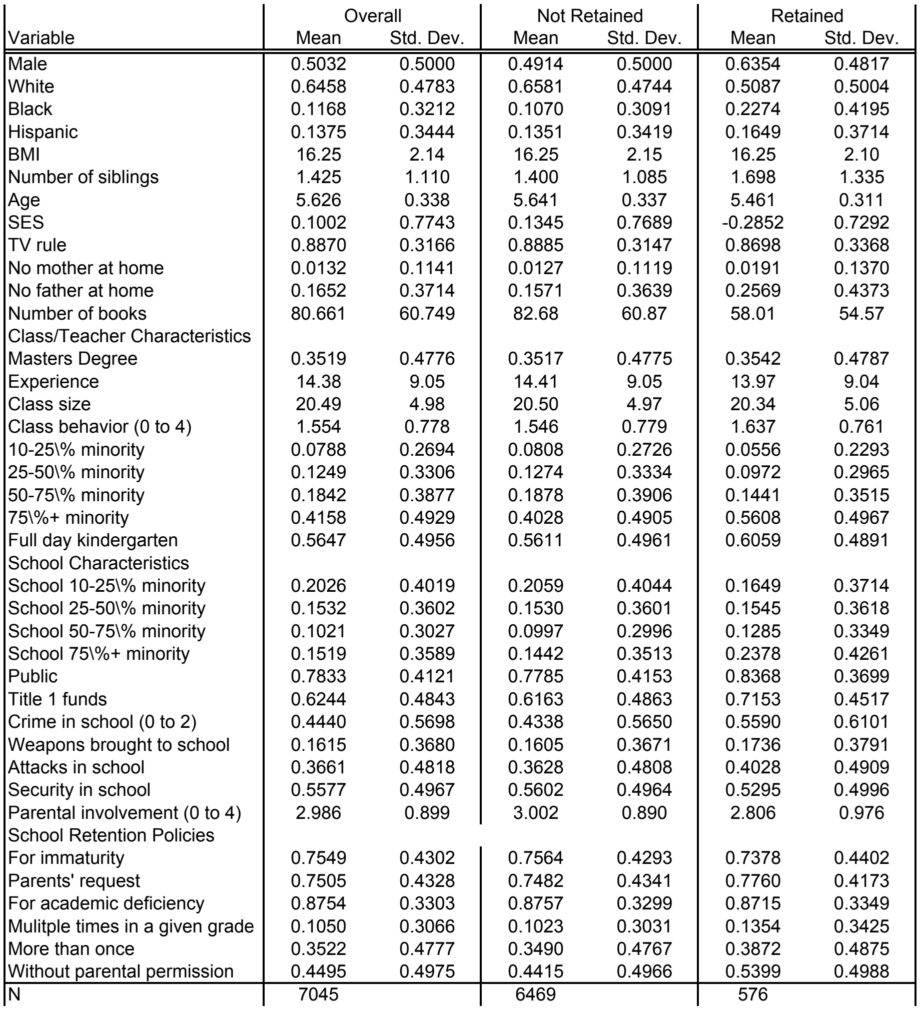
Robin, J.-M. & Bonhomme, S. (2008), Generalized nonparametric deconvolution with an application to earnings dynamics. Unpublished Manuscript.

Schennach, S. M. (2004), 'Estimation of nonlinear models with measurement error', *Econometrica* **72**(1), 33–75.

A Tables and Figures

A.1 Tables

Table 2: Summary Statistics 1998/99



	Overall		Not Retained		Retained	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Male	0.5032	0.5000	0.4914	0.5000	0.6354	0.4817
White	0.6458	0.4783	0.6581	0.4744	0.5087	0.5004
Black	0.1168	0.3212	0.1070	0.3091	0.2274	0.4195
Hispanic	0.1375	0.3444	0.1351	0.3419	0.1649	0.3714
BMI	16.25	2.14	16.25	2.15	16.25	2.10
Number of siblings	1.425	1.110	1.400	1.085	1.698	1.335
Age	5.626	0.338	5.641	0.337	5.461	0.311
SES	0.1002	0.7743	0.1345	0.7689	-0.2852	0.7292
TV rule	0.8870	0.3166	0.8885	0.3147	0.8698	0.3368
No mother at home	0.0132	0.1141	0.0127	0.1119	0.0191	0.1370
No father at home	0.1652	0.3714	0.1571	0.3639	0.2569	0.4373
Number of books	80.661	60.749	82.68	60.87	58.01	54.57
Class/Teacher Characteristics						
Masters Degree	0.3519	0.4776	0.3517	0.4775	0.3542	0.4787
Experience	14.38	9.05	14.41	9.05	13.97	9.04
Class size	20.49	4.98	20.50	4.97	20.34	5.06
Class behavior (0 to 4)	1.554	0.778	1.546	0.779	1.637	0.761
10-25\% minority	0.0788	0.2694	0.0808	0.2726	0.0556	0.2293
25-50\% minority	0.1249	0.3306	0.1274	0.3334	0.0972	0.2965
50-75\% minority	0.1842	0.3877	0.1878	0.3906	0.1441	0.3515
75\%+ minority	0.4158	0.4929	0.4028	0.4905	0.5608	0.4967
Full day kindergarten	0.5647	0.4956	0.5611	0.4961	0.6059	0.4891
School Characteristics						
School 10-25\% minority	0.2026	0.4019	0.2059	0.4044	0.1649	0.3714
School 25-50\% minority	0.1532	0.3602	0.1530	0.3601	0.1545	0.3618
School 50-75\% minority	0.1021	0.3027	0.0997	0.2996	0.1285	0.3349
School 75\%+ minority	0.1519	0.3589	0.1442	0.3513	0.2378	0.4261
Public	0.7833	0.4121	0.7785	0.4153	0.8368	0.3699
Title 1 funds	0.6244	0.4843	0.6163	0.4863	0.7153	0.4517
Crime in school (0 to 2)	0.4440	0.5698	0.4338	0.5650	0.5590	0.6101
Weapons brought to school	0.1615	0.3680	0.1605	0.3671	0.1736	0.3791
Attacks in school	0.3661	0.4818	0.3628	0.4808	0.4028	0.4909
Security in school	0.5577	0.4967	0.5602	0.4964	0.5295	0.4996
Parental involvement (0 to 4)	2.986	0.899	3.002	0.890	2.806	0.976
School Retention Policies						
For immaturity	0.7549	0.4302	0.7564	0.4293	0.7378	0.4402
Parents' request	0.7505	0.4328	0.7482	0.4341	0.7760	0.4173
For academic deficiency	0.8754	0.3303	0.8757	0.3299	0.8715	0.3349
Mulitple times in a given grade	0.1050	0.3066	0.1023	0.3031	0.1354	0.3425
More than once	0.3522	0.4777	0.3490	0.4767	0.3872	0.4875
Without parental permission	0.4495	0.4975	0.4415	0.4966	0.5399	0.4988
N	7045		6469		576	

Table 3: Average Reading Achievement by Retention Status and Year

	1998/99	1999/00	2001/02	2003/04
	(Kindergarten)	(1st grade)	(3rd grade)	(5th grade)
Retention Status:				
Not Retained	3.386	4.305	4.823	4.983
N	7,014	4,267	2,629	1,852
Retained in Kindergarten	3.132	3.773	4.585	4.764
N	246	139	91	76
Retained early	3.076	3.865	4.397	4.671
N	266	153	77	89
Retained late	3.148	3.985	4.524	4.789
N	82	54	31	23

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

Achievement is measured as logs of IRT scores.

Grade headings indicate the grade the student would be in given no retention

Table 4: Average Math Achievement by Retention Status and Year

	1998/99	1999/00	2001/02	2003/04
	(Kindergarten)	(1st grade)	(3rd grade)	(5th grade)
Retention Status:				
Not Retained	3.136	4.091	4.571	4.775
N	7,172	4,302	2,631	1,857
Retained in Kindergarten	2.767	3.638	4.312	4.557
N	252	141	95	74
Retained early	2.671	3.672	4.156	4.441
N	284	160	78	90
Retained late	2.744	3.737	4.244	4.540
N	86	56	31	22

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

Achievement is measured as logs of IRT scores.

Grade headings indicate the grade the student would be in given no retention

Table 5: Social Rating Scale Score: Approaches to Learning
by Retention Status and Year

	1998/99	1999/00	2001/02
	(Kindergarten)	(1st grade)	(3rd grade)
Retention Status:			
Not Retained	0.118	0.152	0.156
N	7,200	4,286	2,617
Retained in Kindergarten	-0.722	-0.092	-0.164
N	255	140	94
Retained early	-0.908	-1.235	-0.703
N	287	158	78
Retained late	-0.402	-0.785	-1.023
N	87	56	31

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

Behavioral test are measured on the Social Rating Scale and standardized with mean 0 and standard deviation 1.

Grade headings indicate the grade the student would be in given no retention.

Table 6: Social Rating Scale Score: Self Control
by Retention Status and Year

	1998/99	1999/00	2001/02
	(Kindergarten)	(1st grade)	(3rd grade)
Retention Status:			
Not Retained	0.060	0.087	0.103
N	7,180	4,267	2,607
Retained in Kindergarten	-0.312	-0.117	-0.197
N	254	140	94
Retained early	-0.407	-0.544	-0.364
N	287	153	78
Retained late	-0.092	-0.218	-0.534
N	87	55	31

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

Behavioral test are measured on the Social Rating Scale and standardized with mean 0 and standard deviation 1.

Grade headings indicate the grade the student would be in given no retention.

Table 7: Social Rating Scale Score: Interpersonal Sklls
by Retention Status and Year

	1998/99	1999/00	2001/02
	(Kindergarten)	(1st grade)	(3rd grade)
Retention Status:			
Not Retained	0.118	0.152	0.156
N	7,200	4,286	2,617
Retained in Kindergarten	-0.722	-0.092	-0.164
N	255	140	94
Retained early	-0.908	-1.235	-0.703
N	287	158	78
Retained late	-0.402	-0.785	-1.023
N	87	56	31

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

Behavioral test are measured on the Social Rating Scale and standardized with mean 0 and standard deviation 1.

Grade headings indicate the grade the student would be in given no retention.

Table 8: Evidence for Dynamic Selection and Treatment Effect (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
Child'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 Measures	--	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

* 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", respectively.

Note: If the p value is small compared to the critical value, we reject the hypothesis of equality of coefficients. P values less than 0.05 are colored with yellow. Yes/No indicates if each group of variables is included as controls.

Table 9: Evidence for Dynamic Selection and Treatment Effect (Math Score)

Dependent Variable	Kindergarten Math Score[#]	Math 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
Child'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 Measures	--	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", respectively.

Note: If the p value is small compared to the critical value, we reject the hypothesis of equality of coefficients. P values less than 0.05 are colored with yellow. Yes/No indicates if each group of variables is included as controls.

Table 10: Predicted and Actual Means and Standard Deviations of Kindergarten (1998-99 School Year) Test Scores/Measures

Test / Measure	Predicted Mean	Predicted Standard Deviation	Actual Mean	Actual Standard Deviation
Reading Test	3.366	0.289	3.364	0.281
Math Test	3.110	0.361	3.103	0.362
Approach to Learning	0.052	0.963	0.047	0.976
Self-Control	0.025	0.964	0.029	0.971
Interpersonal Skills	0.021	0.961	0.018	0.976

Note: Behavioral measures are standardized to have mean zero and variance equal to one. These figures are calculated from 500000 simulations based on the estimated model.

Table 11: Predicted and Actual Retention Probabilities (Conditional on Survival)*

	Data	Model	
		Predicted	Standard Error
Retained in Kindergarten	3.326%	3.519%	
Retained Early (1st or 2nd grade)	4.012%	4.175%	
Retained Late (3rd or 4th grade)	1.325%	1.329%	

Note: The table caclulates the probability of retention at t, conditional on not having been retained before t. Standard Errors obtained via 200 bootstrap replications.

Table 14: Average Reading Test Score Gain by Retention Status: 2003-04 School Year

Average Gain	A kid 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	0.0508	-0.0399	-0.0673	-0.0424	0.0417
Retained Early vs Not Retained	0.0314	-0.0723	-0.0734	-0.0870	0.0221
Retained Late vs Not Retained	0.0596	0.0214	0.0225	0.0198	0.0562

* Statistically different from zero at 5% level

Note: Let T = 0,1,2, or 3 represent the actual retention status of a kid: never retained, retained in kindergarten, retained early (at grade 1 or 2), or retained late (at grade 3 or 4), respectively. Let S(i) be the potential test score if the kid were retained at time i=0,1,2,3. The row i, column j element of this table calculates E[S(i) - S(0) | D=j]. For example, the test score of a kid who was actually not retained would increase by 0.034 if he were retained at 1 or 2 grade instead. When calculating these figures, we keep kid's age fixed at 11.

Table 15: Average Math Test Score Gain by Retention Status: 2003-04 School Year

Average Gain	A kid 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	0.1482	-0.0382	-0.0846	-0.0576	0.1295
Retained Early vs Not Retained	0.0624	-0.0356	-0.0493	-0.0420	0.0530
Retained Late vs Not Retained	0.1599	-0.0142	-0.0447	-0.0258	0.1430

* Statistically different from zero at 5% level

Note: Let T = 0,1,2, or 3 represent the actual retention status of a kid: never retained, retained in kindergarten, retained early (at grade 1 or 2), or retained late (at grade 3 or 4), respectively. Let S(i) be the potential test score if the kid were retained at time i=0,1,2,3. The row i, column j element of this table calculates E[S(i) - S(0) | D=j]. For example, the test score of a kid who was actually not retained would increase by 0.166 if he were retained at 1 or 2 grade instead. When calculating these figures, we keep kid's age fixed at 11.

Figure 1: Density of General Ability

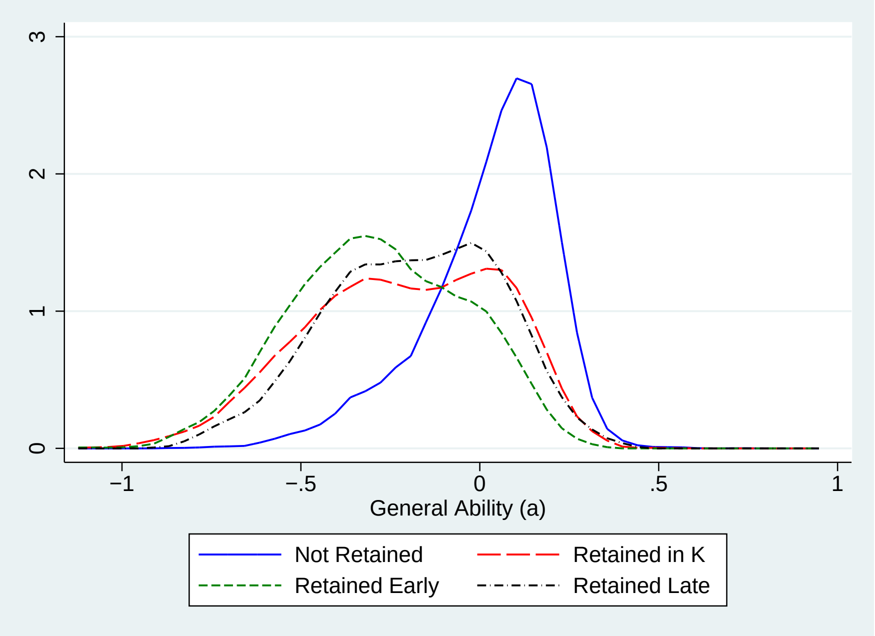


Figure 2: Density of Behavioral Ability

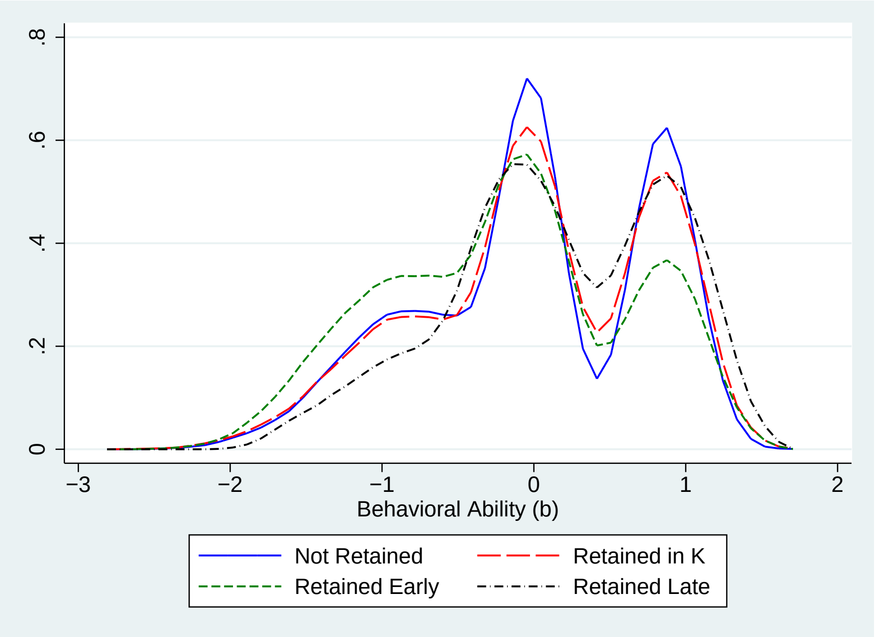


Figure 3: Density of Cognitive Ability

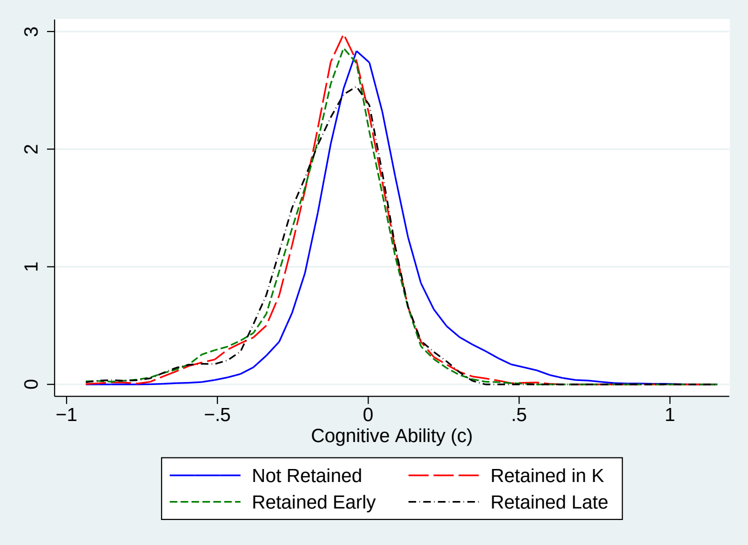


Figure 4: Achievement Gains for Kindergarten Retention over Time

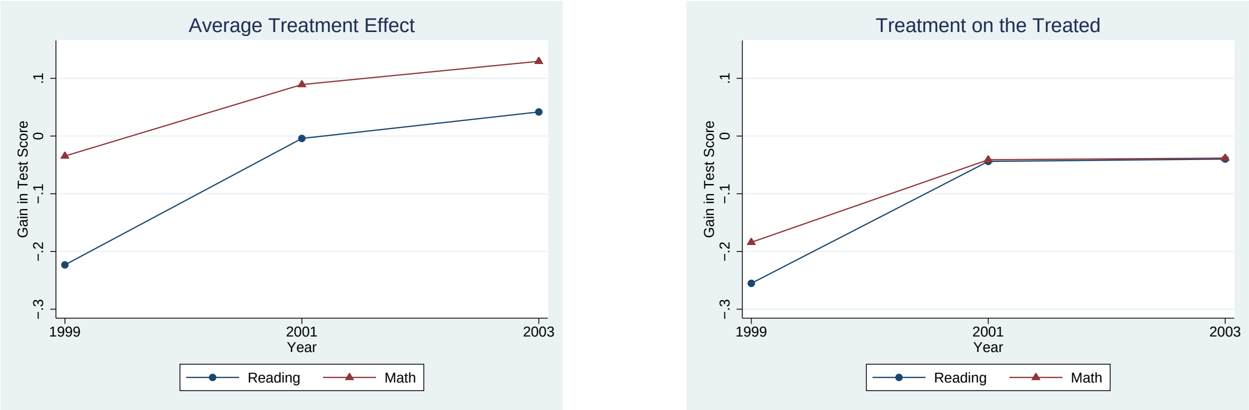


Figure 5: Achievement Gains for Early Retention over Time

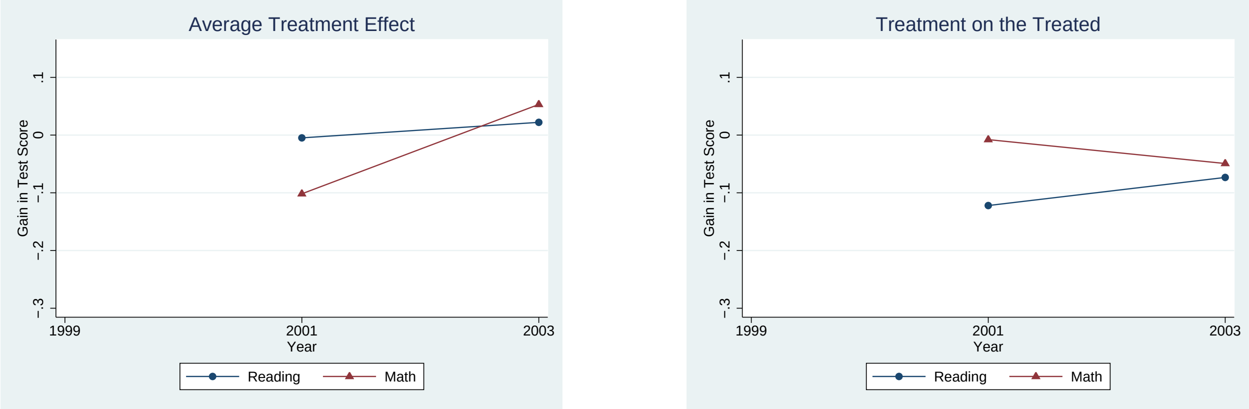


Figure 6: Achievement Gains in 2003/04 by Ability Quantiles

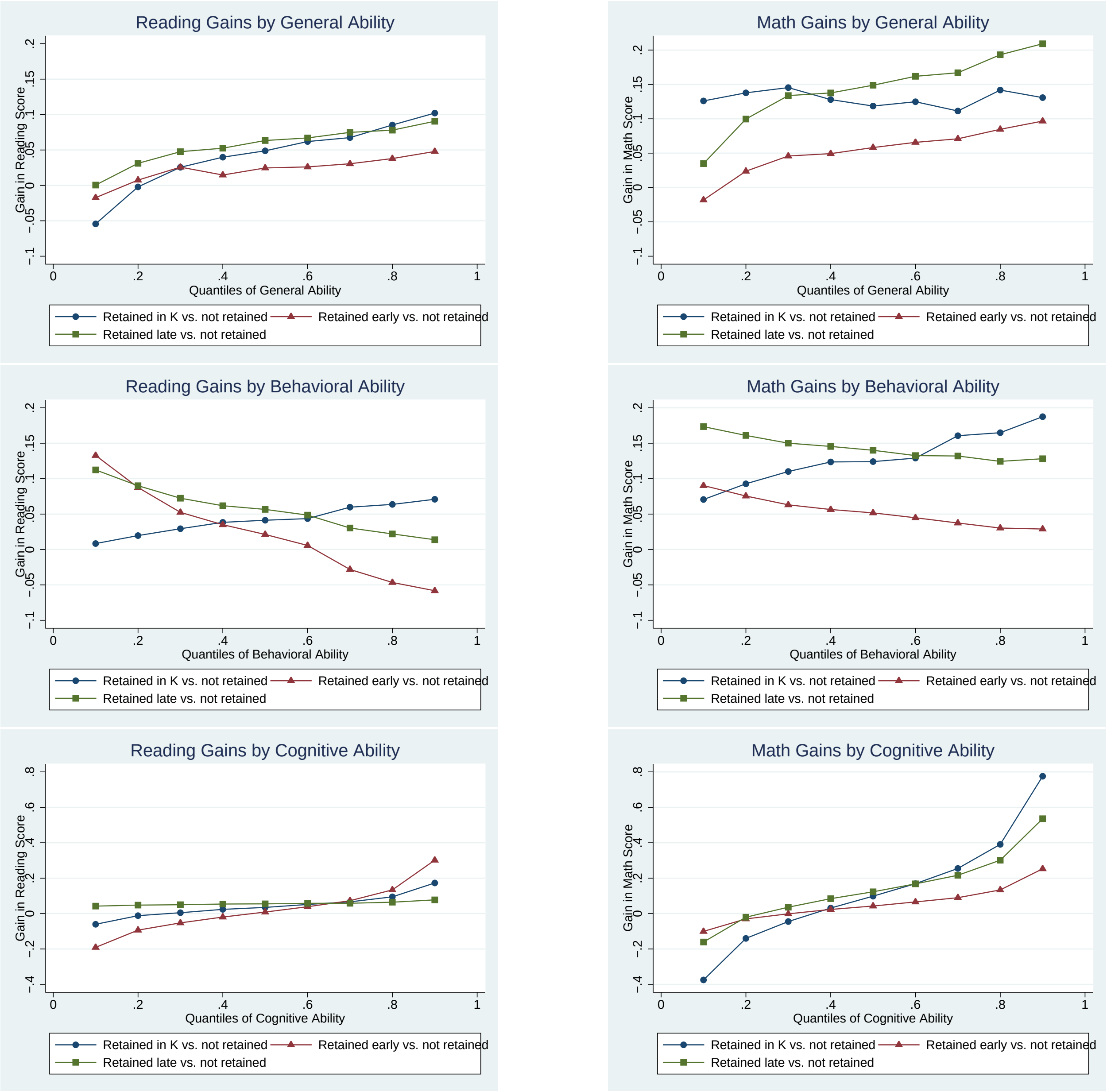
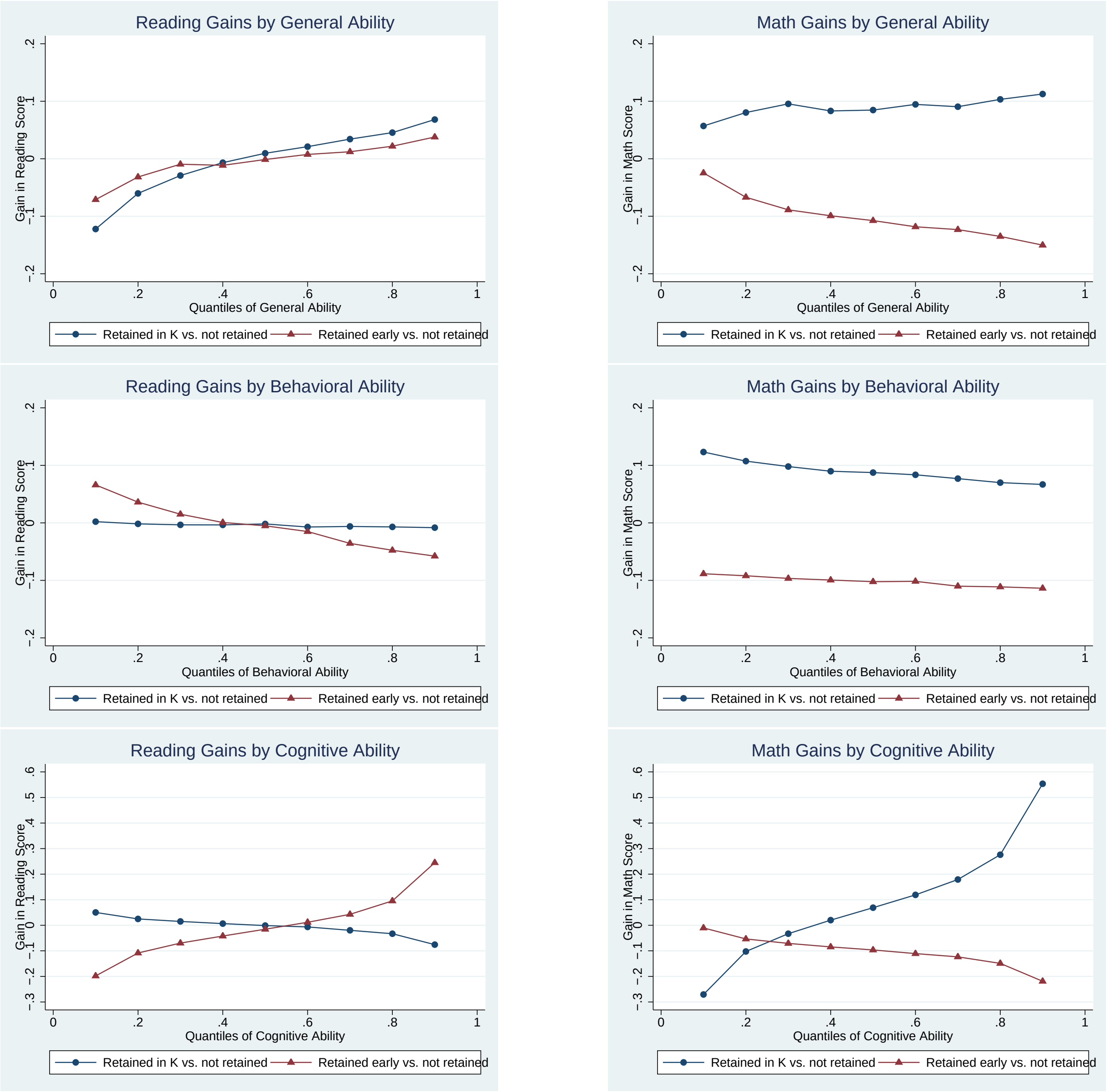


Figure 7: Achievement Gains in 2001/02 by Ability Quantiles



Reading Gains by General Ability

Math Gains by General Ability

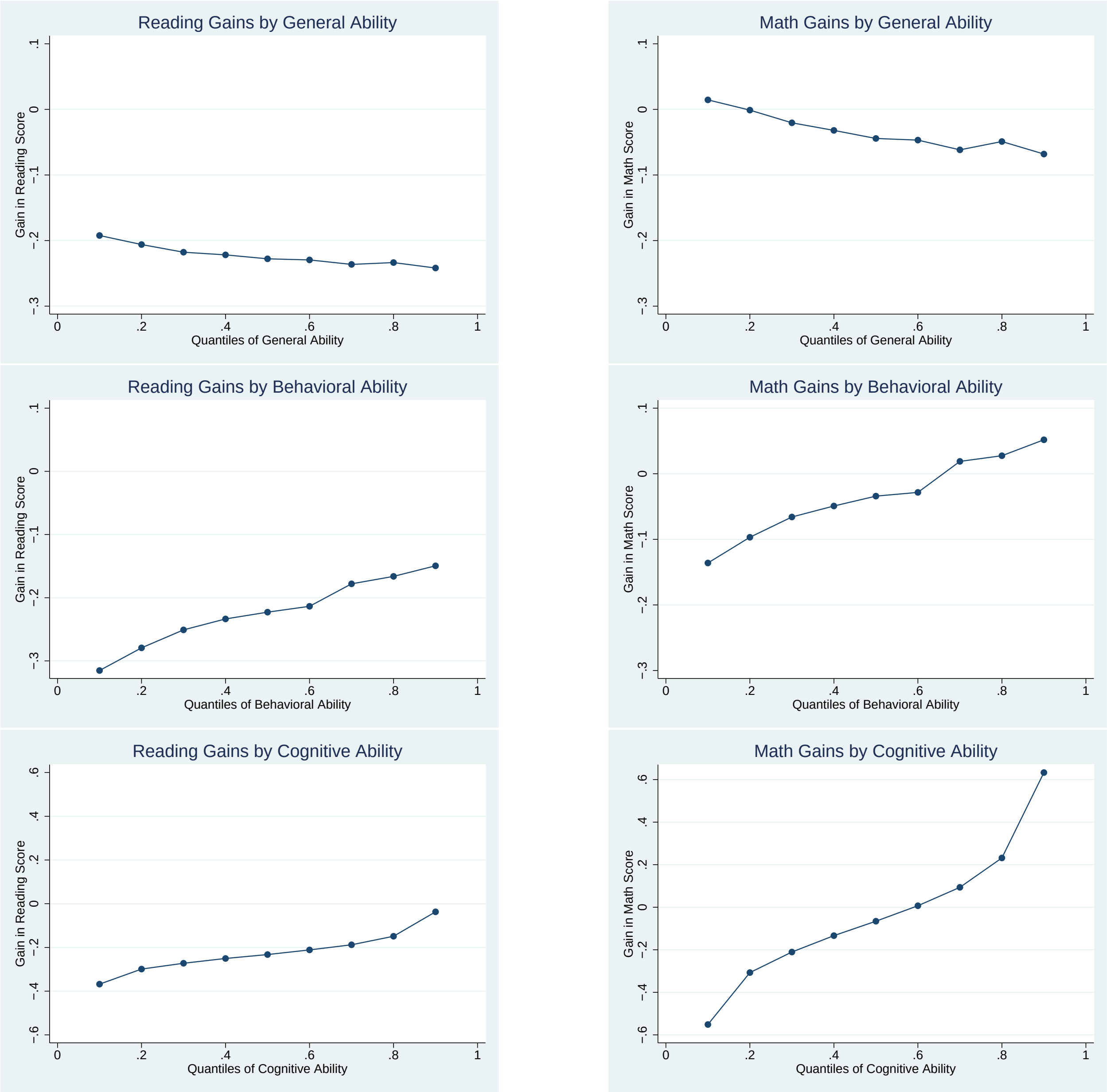
Reading Gains by Behavioral Ability

Math Gains by Behavioral Ability

Reading Gains by Cognitive Ability

Math Gains by Cognitive Ability

Figure 8: Achievement Gains for Kindergarten Retention in 1998/99 by Ability Quantiles


Reading Gains by General Ability
Gain in Reading Score
Quantiles of General Ability
Math Gains by General Ability
Gain in Math Score
Quantiles of General Ability
Reading Gains by Behavioral Ability
Gain in Reading Score
Quantiles of Behavioral Ability
Math Gains by Behavioral Ability
Gain in Math Score
Quantiles of Behavioral Ability
Reading Gains by Cognitive Ability
Gain in Reading Score
Quantiles of Cognitive Ability
Math Gains by Cognitive Ability
Gain in Math Score
Quantiles of Cognitive Ability