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

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

May 28, 2011

Abstract

Increasingly, grade retention is viewed as an important alternative to social promotion, yet evidence to date is unable to disentangle how the effect of grade retention varies by abilities and over time. The key challenge is differential selection of students into retention across grades and by abilities. Because existing quasi-experimental methods cannot address this question, we develop a new strategy that is a hybrid between a control function and a generalization of the fixed effects approach. Applying our method to nationally-representative, longitudinal data, we find evidence of dynamic selection into retention and that the treatment effect of retention varies considerably across grades and unobservable abilities of students. Our strategy can be applied more broadly to many time-varying or multiple treatment settings.

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

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

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

[‡]Department of Economics, University of Western Ontario. email: s.navarro.lozano@gmail.com

[§]Department of Economics, University of Mannheim. email: ytakahas@mail.uni-mannheim.de

1 Introduction

Grade retention (or grade repetition) is a common practice in many countries. In the U.S. about 10 percent of students are retained between kindergarten and eighth grade.[1] With states facing increasing pressure to ensure that students meet minimum proficiency levels under the No Child Left Behind Act (NCLB) of 2001 and associated national rhetoric emphasizing an end to social promotion,[2] the practice of grade retention is facing new scrutiny as a potential policy to help bring students up to proficiency levels.[3] This is somewhat surprising given that the empirical literature at best provides mixed evidence of its effectiveness in improving student outcomes (e.g., Holmes, 1989; Jimerson, 2001).

In this paper we provide evidence on how the effect of grade retention varies by age and unobserved abilities based on a model for multiple/time-varying treatment effects. The key challenges we address are that a) grade retention is not a binary treatment problem (i.e., a child can be retained in one of multiple grades), b) the effect of grade retention is likely to vary by students' unobservable behavioral and cognitives abilities and c) the effect is also likely to vary depending on the time at which a student is retained and the time elapsed since retention. For the most part, the literature has not addressed these three issues. For the most part the literature has focused on attempting to address the problem that retained children are different from non-retained ones (i.e., standard selection). This is an issue we also address, but with the non-trivial extension to a multiple treatment/time-varying setting and the complications arising from it.

Heterogeneity in responses to grade retention by abilities and by timing is central to informing policy. For instance, student accountability policies based on retention may vary in effectiveness depending on the average ability of the students who are retained (as determined by the cutoff for passing the exam) and whether the policy applies to students in early or later grades. This may be particularly important given evidence of increasing retention at early ages (Hauser, Frederick and Andrew, 2007). Furthermore, one of the primary criticisms of grade retention is that it will disrupt a student's social connections and affect their socio-emotional development. The negative social consequences of being retained in kindergarten are may be less severe than retaining a child in her teenage years. We provide new evidence on the importance of these sources of heterogeneity in understanding grade retention. As such, our findings also shed light on some of the mixed evidence in the literature.

[1] See National Center for Education Statistic, Condition of Education 2009, Indicator 18, http://nces.ed.gov/programs/coe/2009/section3/indicator18.asp.

[2] Hauser, Frederick and Andrew (2007) find an increasing trend in grade retention since 1996, with some sharp increases at the time of NCLB for some states.

[3] Eighteen states specify particular assessments to be used for grade promotion. Twelve states specify promotion gateway grades, a performance threshold for promotion to the next grade (Zinth, 2005).

We develop a simple framework to address how the timing of grade retention affects student outcomes and how the effects vary by students abilities. Our method can be understood as a hybrid between a control function and a generalization of the fixed effect approach. We assume that a "low" dimensional set of unobservables affects both selection into treatment and the outcome of treatment. This strategy effectively places restrictions on the covariances between unobservables in the outcome and selection equations, a generalization of the semiparametric factor structure of Carneiro, Hansen and Heckman (2003).[4] It is a control function approach because we use information from the selection equation to help control for selection, so that the same unobserved abilities affect both test scores and the probability of being retained. Identification is further enhanced by the use of exclusion restrictions (retention policy variables), variables that affect a student's selection into retention that do not affect their outcomes directly.

Existing methods are not well-equipped to study these dimensions of grade retention. One branch of the literature controls for selection based only on *observables* through matching methods.[5] These papers attempt to correct for selection by creating a control group of non-retained students based on observable characteristics, such as IQ, measures of socio-emotional adjustment, academic achievement, socio-economic status (SES) and gender. However, grade retention is a particularly compelling case where even with a rich set of controls, retained students are likely to have unobservable characteristics that lead to their retention relative to observationally similar students who were not retained. These unobservables may help explain why studies based on matching generally find negative effects of grade retention, i.e., the control group may be "better" in unobservable dimensions. Furthermore, extending matching on observables to consider time-varying effects of grade retention requires even stronger assumptions: that conditional on observables, the probability of being retained is (sequentially) independent over time.[6] As both the assumption of selection on observables and the repeated conditional independence assumptions seem particularly problematic for our context, we focus on the case of selection on unobservables. More precisely, we are interested in the case where the gains from retention may vary based on the unobservable "abilities" of the student and these unobservable gains determine retention (*essential heterogeneity* in the language of Heckman, Urzua and Vytlacil (2006)).

Only a few studies of grade retention attempt to control for selection on unobservables,

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

[5]See Jimerson (2001) for overview.

[6]See Gill and Robins (2001), Murphy (2003) and Lechner (2004) for recent developments. See also Frölich (2004) and Cattaneo (2010) for multivariate static treatment models that invoke similar conditional independence assumptions.

and these generally present a more positive picture of the effects of grade retention. Jacob and Lefgren (2004) and Jacob and Lefgren (2009) use a regression discontinuity approach that compares students just above to students just below a threshold for passing imposed by Chicago's student accountability policy. This provides an important contribution to the literature and is a useful approach for estimating a *local* effect of grade retention (the effect for marginal students in a particular grade relative to a particular policy in place). However, it cannot inform our question of interest: whether the effect varies for students well below the threshold, which may be more of a target group of the policy. Furthermore, it cannot be extended to study variation in effectiveness across grades in our context of heterogeneous treatment effects.[7]

Comparing the estimated effect of grade retention across grades using a local approach could lead to erroneous conclusions about whether it is better to retain a child earlier or later for at least two reasons. First, different types of students may select into retention at different ages; for instance, if a student is already retained in kindergarten, that same student may be much less likely to be retained in first grade, so the pool of potential retainees varies over time. We term this *dynamic selection.* Second, the difficulty in meeting the threshold for passing may vary across grades, so that the marginal student who is retained in one grade may differ from the marginal retained student in another grade. This means that comparing the estimated effect of grade retention across grades using regression discontinuity confounds selection into retention with potentially heterogeneous effects of retention. Similar problems arise with extending the instrumental variable approach applied by Greene and Winters (2007), which identifies the effects of grade retention by exploiting the introduction of Florida's test-based promotion policy. Furthermore, the existing quasi-experimental findings using these threshold selection rules may not generalize to the majority of settings where the selection rule is less clear and different types of students may be retained. For instance, in some cases students may be retained through parental request, or because of low behavioral ability rather than underperformance on a cognitive exam.

Given that existing methods are not well-equipped to answer the questions we pose, we develop a new strategy with several objectives in mind. First, a key feature of our method is to control for dynamic selection into grade retention, i.e., that the unobservable types of students who select into retention are likely to vary across grades, as we find that this is an important phenomenon in the data. Second, our method recognizes that the effect of grade retention is likely to differ by the unobservable abilities of students. Third, our method provides a rich representation of the unobservable abilities that may affect selection into

[7]Cellini, Ferreira and Rothstein (2010) provide a generalization to the regression discontinuity approach to time-varying treatment effects in the context of homogeneous treatment effects.

and the outcome of retention, particularly taking account of both cognitive and behavioral abilities along with time-varying shocks to achievement. Finally, our method is sufficiently general to extend to the predominant setting where the selection rule is unknown.

We evaluate the effect of retention on achievement using data from the Early Childhood Longitudinal Study of Kindergartners (ECLS-K). We find that students who are retained in kindergarten would have performed as much as 27 percent higher in the next year if they had not been retained. We also find that the initial losses to achievement diminish over time. By the end of our data, when students are approximately age 11, eliminating grade retention increases raises achievement by as much as 7 percent for students who were retained in prior years. This means that these retained students learn 7 percent less by age 11, than they would have learned if they had not been retained. As we discuss further below, a somewhat surprising finding is that the treatment effect of kindergarten retention is positive for the average untreated student in the long run, whereas it is negative for the average treated student. . In comparison, the simpler fixed effect approach only provides, at best, estimates of the average treatment effect. Since the average student is higher ability than the typical student retained, the fixed effect approach would lead to erroneous policy conclusions.

The contribution of our paper extends beyond providing new evidence on the effect of grade retention. Arguably, our method is useful for a broader array of applications where timing matters or there are multiple potential treatments and selection on unobservables is likely to be important. Abbring and Van den Berg (2003) and Ham and LaLonde (1996) provide other useful approaches to analyzing treatment effects in dynamic models. Whereas they rely on the proportional hazards assumption, our model supports more general forms of treatment heterogeneity than in either Ham and LaLonde (1996) (where treatment effects are homogeneous), or Abbring and Van den Berg (2003) (where treatment heterogeneity can be allowed at the expense of ruling out the endogenously-selected time-at-treatment to affect outcomes). Our approach to modeling time-varying treatments is close to that in Heckman and Navarro (2007). However, our focus is substantively different, i.e., how factor analytic methods can aid in identification and interpretation of time-varying treatment effects. We provide a further methodological contribution in generalizing the factor structure results used in other settings (Carneiro, Hansen and Heckman, 2003; Bonhomme and Robin, 2010). Our generalization is appealing not only because it is likely to be useful in other settings, but also because but also because we require a smaller number of measurements (requiring less data) to recover the factors than in existing factor models.[8]

[8]Furthermore, we link the assumptions used in factor structure models to better known fixed effects and regression discontinuity approaches. In online Appendix C we provide further comparisons with other commonly employed methods.

The paper proceeds as follows. In Section 2, we describe the basic framework and define dynamic treatment effects for the dynamic case. In Section 4, we specialize the framework to our proposed factor structure. We show that the model is semiparametrically identified. We describe our estimation strategy in Section 5. Data and results are discussed in Sections 3 and 6.

2 The Framework

As discussed in the introduction, existing methods for estimating static treatment effects that have been applied to the study of grade retention do not extend readily to the context where the treatment effect of retention varies across grades and by student abilities. Below we outline a simple framework for evaluating the effect of grade retention in a general context where the selection rule is unknown. While our generalizations come at the expense of considerable notation, the framework provides important insight into the challenges associated with estimating time-varying treatment effects that vary by ability.

Let $t = 1, 2, ..., \bar{t}$ index calendar time and $i = 1, ..., I$ index the individual. Since we allow for students to be retained at different times, we define a random variable that indicates the grade in which a student is retained, $R_i = \{1, 2, ..., \bar{R} - 1, \bar{R}, \infty\}$, where $\bar{R} \leq \bar{t}$ allows for the possibility that students may be retained only up to a certain time period or grade.[9] We assume that the student is retained at most once.[10] Our data follows a single cohort of kindergarteners across time, so that $R_i = 1$ denotes that a student is retained in kindergarten, etc. We adopt the convention of letting $R_i = \infty$ for the "never" treated state where a student is not retained.[11]

The (possibly vector-valued) outcome of interest, math and reading test scores in our context, at time t for a student i who receives treatment at time r is denoted by $Y_i(t, r)$.[12] For notational simplicity, we keep all conditioning on covariates, observable school and student characteristics, implicit. Finally, we define a random variable $D_i(r)$ that takes value 1 if an individual is retained at time r and 0 otherwise. For individual i the observed outcome in

[9] Note that we could restrict this further by allowing a lower bound on when students could be retained, but it does not apply in our data.

[10] We do this primarily because only 0.3 percent of students are retained twice in our data and this assumption simplifies notation. However, extending the framework to allow for the possibility of treatment being taken more than once is straightforward, by letting R_i be a random vector characterizing the times at which an individual receives treatment.

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

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

period t will be given by

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

As opposed to the standard binary treatment case, we now have many potential outcomes. That is, while the standard case only has the treated and untreated potential states, we have the untreated, the treated at time 1, the treated at time 2, etc. Because of the sequential nature of the problem, by letting $Y_i(t,r)$ depend on treatment time r, we allow for the possibility that the effect of treatment depends not only on receipt but on the time at which treatment is received. That is, there is no single effect of retention, but rather an effect of retention in kindergarten, in first grade, etc. Furthermore, there is no single effect of retention in kindergarten (for example), as the effects depend on the time elapsed since retention.[13] We define the set of potential treatment effects below.

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

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

This assumption rules out that potential outcomes differ because in the *future* treatment times will be different. This means, for example, that after conditioning on all prior information, the fact that a student will be retained in second grade does not directly affect her performance in first grade. While Abbring and Van den Berg refer to this as the *no anticipations* assumption, this should not be confused with the assumption that individuals are not forward looking. Assumption **A-1** does not rule out that individuals may predict that they are more likely to get treated at a particular time r (i.e., have some anticipation as to treatment time).[14]

We further write the outcomes as

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

where, because of **A-1**, we impose $\Phi(t,r) = \Phi(t)$ and $\epsilon_i(t,r) = \epsilon_i(t)$ if $r \geq t$.[15] The

[13] This setting can also be interpreted as depending on the time since treatment $(t-r)$, making it straightforward to analyze the outcomes as durations, counts, etc.

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

[15] While we focus on continuous test scores in our application, we can easily use this framework to 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

observed outcome (test score) in period t is then given by

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

In most cases, the decision to retain a student is not clearly defined, but rather is the result of a complex process involving many actors, including teachers, principals and parents. We thus model selection in a reduced form way that still highlights the importance of timing, such that treatment and treatment time are determined by a single spell duration model that follows a sequential threshold crossing structure as in Heckman and Navarro (2007). If we define the treatment time specific index $V_i(r) = \lambda(r) + U_i(r)$ for $r \in \{1, 2, ..., \bar{R}-1, \bar{R}\}$, then treatment time is selected according to

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

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

This framework can be thought of as a midpoint between the standard static treatment literature that does not model the selection process explicitly and a fully specified structural dynamic discrete choice model.[16] At the same time, the selection process we propose is consistent with, for example, the commonly employed test score thresholds for whether a child should repeat a grade. This threshold could be individual specific if schools use relative comparisons or take into account extenuating circumstances such as being a special education student. Notice that our selection process applies whether we observe the scores used for the decision or not. For example, if the j^{th} test score $Y_{i,j}(t)$ (whether observed by the

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

[16] Our selection model is consistent with the usual threshold-crossing or reservation-value decision rules that frequently arise from complex dynamic decision problems. Cunha, Heckman and Navarro (2007) provide conditions under which structural dynamic discrete choice models can be represented by a reduced form approximation as above. Furthermore, since extending it to the case in which treatment is not an absorbing state (i.e., treatment can be received more than once) is straightforward, it can be applied in more complex situations. In this case, we would generalize the threshold crossing model into a multiple spell model, where the whole sequence of prior treatments/no treatments potentially affects the decision each period. R_i would be a vector containing the treatment history up to t, and an individual would choose treatment every time the index becomes positive (not only the first time).

econometrician or not) is used to decide who to retain, and the threshold μ_i is individual specific, we would have

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

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

2.1 Defining Treatment Effects

As mentioned above, our framework encompasses many potential treatment effects because both the timing of treatment and the time elapsed since treatment may matter. Thus, before turning to the identification problem, we first consider the problem of defining what constitutes "the" effect of treatment at the individual level. A particular parameter of interest for the individual treatment effect

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

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

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

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

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

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

etc. For instance, we may want to know the return to retaining students in kindergarten

who were actually retained in first grade.[17]

3 Data

We use the ECLS-K, a nationally representative survey of kindergartners in 1998/99, to study the effect of grade retention. It follows the students as they progress through school, with follow-up surveys in the 1999/2000, 2001/02 and 2003/04 school years. A benefit of these data is that we observe the history of a student's schooling beginning at kindergarten, and it covers the earlier years when retention is relatively more common. Roughly 10% of our sample is retained between kindergarten and fourth grade. We restrict the sample to students who were retained only once, did not skip grades, and were taking kindergarten for the first time in 1998/99.[18] Because of the nature of the survey, we are able to form three different retention indicators: kindergarten, early (first or second grades) and late (third or fourth grades).[19] That is, our dynamic treatment time indicator takes values $R_i = 1, 2, 3, \infty$, where $R_i = \infty$ means the child is never retained, $R_i = 1$ that he is retained in kindergarten, $R_i = 2$ that he is retained early and $R_i = 3$ that he is retained late.

Each year of the ECLS-K includes cognitive tests measuring students' science, reading and math skills.[20] We focus primarily on the effect of retention at different grades on the math and reading tests, using the log of the item response theory (IRT) scores. ECLS-K also includes measures of teacher ratings on students' behavioral and social skills—the approaches to learning, self-control and interpersonal skills components of the Social Rating Scale (SRS). We use these together with the cognitive tests in order to identify the different components of ability as described below in Section 4.

A logical difficulty in evaluating the effect of grade retention is that it is impossible to hold both the grade and age fixed when determining the gains in achievement for a retained student. Depending on the policy question of interest, it may be more appropriate to focus

[17]Under certain assumptions that limit the heterogeneity of treatment effects some of these parameters may equal one another by construction. We focus on the more general case, where the treatment effect is allowed to vary over time and by unobserved individual characteristics. Both of these types of heterogeneity prove important in our application.

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

[19]In principle we could separate early and late into the four grades at which retention takes place. This, however, can only be done for less than half of the sample, and we already lose a significant amount of data because of attrition, as shown below.

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

on measuring effects holding grade fixed or holding age fixed. The effect holding grade fixed addresses, for instance, whether a student learns more by the end of fifth grade than he would have if he had not repeated fourth grade. This attributes maturation (or age) effects to the estimated effect of retention. Alternatively, holding age fixed measures whether a student learns more, say, by age 11 if he 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. That is, the tests used by ECLS-K are designed to measure cognitive development as opposed to grade-specific knowledge.[21]

The ECLS-K contains a very rich set of covariates. We use characteristics of the children, the family, the class and the school as controls in our model. Class and teacher characteristics are taken from teacher surveys.[22] School administrator surveys provide information about the school characteristics, and parent surveys provide information about the family.

Table 1 shows descriptive statistics for the covariates we include in all our equations for the first year of the survey (1998/99) in columns 2 to 4. We restrict the sample to students who have any test score measure in the first year and the full set of conditioning covariates. Thus, the number of observations differs across test scores and covariates. We do this so that we can include as much of the data as possible in estimating the different outcome equations. A potentially important concern with a panel study of this type is non-random sample attrition. Column 6 of Table 1 shows the mean 1998/99 characteristics for students who are still in the sample in 2003/04 (the last year of the survey that we use for estimation). The number of observations decreases substantially across these years, from 7832 in the base year to 2106 in the last year. Comparing summary statistics, we see suggestive evidence of non-random attrition. Our estimator controls for this, as discussed in Section 4.3.

The ECLS-K also includes information on the schools' retention policies for the 1998/99, 1999/00 and 2001/02 survey years. These policies include whether the school has a policy that allows children to be retained in any grade (this policy only applies to grades after kindergarten), to be retained because of immaturity, to be retained at the parents' request, to be retained without parental authorization, to be retained multiple times or multiple times in a given grade. As shown in Table 2, retention policies vary considerably across schools and also to a lesser extent across retention statuses. In general, students who are retained early or late attend schools with more "liberal" retention policies than students who are not

[21]Roughly speaking it is like being given the same test every year and measuring how many additional questions the student can answer.

[22]For the 2003/04 school year, both math/science and reading teachers fill out surveys, resulting in potentially different classroom and teacher characteristics for math/science and reading. We use the relevant classroom measures for each test in estimating the outcome equations.

retained or who are retained in kindergarten. For instance, in the 1998/1999 school year 44% of schools in the non-retained sample permit retention without parental permission, compared to 61% and 58% for students who are retained early or late. Our methodology in Section 5 incorporates these variables by using them as exclusions, under the assumption that, conditional on the other covariates including observable school characteristics, they do not directly determine the child's test score but they do affect the probability that a child repeats a grade.

4 Identification

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

We perform some baseline OLS regressions that indicate that dynamic selection and/or time-varying treatment effects are likely to be important in our data. To test for dynamic selection, we regress the kindergarten cognitive tests, which took place prior to any retention decisions, on period-specific indicators of whether the child is retained in the future. We also control for covariates related to the child, his family, school and class, as described in Table 1 above. Column 2 of Table 3 presents results for reading and math in Panels A and B respectively. Not surprisingly, children who will be retained have lower kindergarten test scores than those who will not be retained. Reading scores are 18% lower for kindergarten retainees, and 20% and 12% lower for early and late retainees. Math scores are even more striking, 27%, 32% and 22% lower for kindergarten, early and late retainees respectively. Furthermore, p-values, reported at the bottom of the table, reject the joint test that the coefficients on being retained at different grades in the future are the same. These results suggest not only the presence of selection but also *dynamic* selection on cognitive test scores in the sense that different types of students are being retained at different grades.

We show evidence that time-varying treatment effects are likely to be present by regressing test scores in the last sample period (2003/04 school year) on retention in different grades. As shown in column 3 of Table 3, being retained is associated with worse outcomes than not being retained. The coefficients on the different retention statuses are also significantly different from each other. This is not direct evidence of time-varying treatment effects, since

differences in the estimated effects across grades could be a result of time-varying treatment effects or a result of dynamic selection.

One way to begin to control for a static component of selection is to include various performance measures in kindergarten, prior to any retention decisions taking place. Columns 4 and 5 present results controlling for kindergarten cognitive test scores and then behavioral scores. Consistent with the existence of selection, the negative effects of retention become smaller but do not disappear. For instance, the coefficient on kindergarten retention is cut in half for both reading and math, from -18% without initial test controls to -9% with test controls. Furthermore, we reject the formal test of equality of the effects for different retention times, again providing evidence for potentially time-varying treatment effects.[23] After including all initial test controls, retention in kindergarten is estimated to lower achievement by 9%, early retention by 14% and late by only 4% in both reading and math.

While this provides suggestive evidence of both time-varying treatment effects and dynamic selection, it is far from conclusive. The assumption that kindergarten test scores control for dynamic selection is a very restrictive one, in that it assumes a static ability that determines whether one is retained in kindergarten, early or late. In addition, tests scores are noisy measures of true latent abilities; hence using the kindergarten measures as controls may actually worsen the bias in the estimated treatment effects.[24] Furthermore, this analysis does not capture heterogeneous effects of treatment by student type, which is a central motivation of our paper.

As we discussed in the introduction, one can also attempt to control for selection on unobservable abilities using regression discontinuity and instrumental variables methods. While these methods are useful for identifying a local treatment effect, they are less well-equipped to determine how the effect of treatment varies by unobservable abilities of students, which may be a key aspect of developing effective grade retention policies. Second, these rely on a particular selection rule, meeting a proficiency threshold on an end of year exam, which does not apply in most settings. Third, because of the local nature of the estimates, they cannot be easily extended to study time-varying treatment effects when there are heterogeneous treatment effects, simply because the student on the margin of being retained varies across grades.[25]

As a consequence, in this section we develop a methodology based on a factor-analytic approach for dealing with dynamic selection and heterogeneous, time-varying treatment effects. We then describe conditions such that the model is semiparametrically identified.

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

[24]See Heckman and Navarro (2004).

[25]In online Appendix C we briefly and more formally discuss some of the advantages and shortcomings of applying commonly employed approaches in the static treatment literature in our time-varying setting.

Our approach can be understood as a hybrid between the control function and a generalized version of the fixed effect approach, as we discuss further below. As with all control function based methods, identification is more transparent and easier to achieve when variables are available that affect selection into treatment but not the effect of treatment. We will use school retention policies as exclusions in our application, but they are not strictly required. In contrast to the standard fixed effect approach, we can allow for the individual effects to be multidimensional, time-varying and treatment-specific (e.g., the effect of ability can differ in the retained relative to the non-retained states).

4.1 Factor Structure

Exploring the important dimensions of timing and heterogeneity in treatment effects comes at the expense of a somewhat more complicated identification strategy than is currently used in the literature. We attempt to simplify exposition by illustrating our strategy with a 3 period example, where treatment can be taken in either of the first 2 periods ($R = 1, 2$), e.g., students can be retained in kindergarten or first grade. The policy is evaluated according to its effect on some outcome measured at period t: $Y_i(t, r)$, e.g., third period test scores. For example, potential outcomes in period 3 can be given by

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

and the observed outcome can be written as

$$\begin{aligned} Y_i(3) &= \Phi(3, \infty) + D_i(1)[\Phi(3, 1) - \Phi(3, \infty)] + D_i(2)[\Phi(3, 2) - \Phi(3, \infty)] \\ &\quad + \epsilon_i(3, \infty) + D_i(1)[\epsilon_i(3, 1) - \epsilon_i(3, \infty)] + D_i(2)[\epsilon_i(3, 2) - \epsilon_i(3, \infty)]. \end{aligned} \quad (4)$$

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

One way to account for essential heterogeneity is to recover the joint distribution of all the unobservables (U_i, ϵ_i). This way we can describe how the treatment effect varies

across unobservable individual types. Imposing a factor structure simplifies the problem and permits us to recover the joint distribution of the unobservables. In particular, we assume:

A-2 *(Factor structure)* $\epsilon_i(t,r) = \theta_i\alpha(t,r) + \varepsilon_i(t)$ and $U_i(r) = \theta_i\rho(r) + \upsilon_i(r)$ where θ_i is a vector of mutually independent "factors" and we assume that $\varepsilon_i(t) \perp\!\!\!\perp \varepsilon_i(t')$ for all $t \neq t'$, $\upsilon_i(r) \perp\!\!\!\perp \upsilon_i(r')$ for all $r \neq r'$ and $\upsilon_i(r) \perp\!\!\!\perp \varepsilon_i(t)$ for all r and t where $\perp\!\!\!\perp$ denotes statistical independence.[26]

We impose **A-2** for convenience, even though it is stronger than required.[27] The factor structure assumption is a convenient dimension reduction technique: it reduces the problem of recovering the entire joint distribution of (U_i, ϵ_i) to that of recovering the factor "loadings" $\alpha(t,r)$ and $\rho(r)$ and the marginal distributions of the elements of θ_i and of $\varepsilon_i(t), \upsilon_i(r)\ \forall t, r$.

The factor structure also has an appealing interpretation, since we can now talk about a low dimensional set of common "causes."[28] The same set of unobservables (the vector θ_i) that determines the effect of grade retention also determines whether a student is retained. We think of θ_i as a vector of unobserved "abilities" in our setting, where essential heterogeneity arises because unobserved abilities affect both the gain in test scores across two years and the probability of being retained. We can then consider questions such as whether less able students in our model are more likely to be retained earlier or later and test the implications for the effect of treatment on these students. Notice that, by writing the selection equation directly as a function of abilities, θ, selection depending on test scores becomes a special case of our choice process as shown in equation (3).

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

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

The observed outcome vectors are

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

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

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

[27] Following the analysis of measurement error models in Schennach (2004) and Hu and Schennach (2008) we can relax the strong statistical independence assumptions and replace them with a combination of general dependence and weaker mean independence assumptions.

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

and

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

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

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

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

If we could recover (or condition on) the unobserved θ_i, then $D_i(1)$ and $D_i(2)$ would no longer be endogenous, and we could obtain consistent estimates of the treatment effect. The factor structure can thus be also understood as an alternative form of matching, where the idea is to "match" based not only on variables observable to the econometrician but also on the unobservable factors. This is the key intuition behind the factor model, to condition not only on observable covariates but also on the unobservable vector θ_i in order to recover the conditional independence assumption of quasi-experimental methods.

To understand how the factor model we propose is a generalization of the fixed effect model, take differences between the period 2 and period 1 j^{th} outcomes to difference out the individual effect θ_i, so that

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

For the differencing strategy to work, we need to impose two restrictions. First, we would need to rule out essential heterogeneity, i.e., $\alpha(2,1) = \alpha(2,\infty) = \alpha(2)$. Second, we would have to assume that the marginal effect of θ_i does not change over time so $\alpha(2) = \alpha(1) = \alpha$. First differencing eliminates θ_i only when these two restrictions hold. As more periods pass, more assumptions are required for the fixed effect model to work. For instance, to identify the effect on period 3 outcomes, we would need to impose the additional assumption that $\alpha(3,2) = \alpha(3,1) = \alpha(3,\infty) = \alpha(3)$.[29]Of course, these assumptions that make the differenc-

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

ing strategy possible, also make the average treatment effect the same as the treatment on the treated. As argued above, this is not a reasonable assumption in the case of grade retention, as in many setting. Thus, our factor structure provides an important generalization of the fixed effect approach by allowing for essential heterogeneity and that the marginal effects of abilities vary by treatment status. In addition, our factor structure further generalizes from the fixed effect approach by permitting multidimensional abilities, so that retention decisions and the outcome of retention can depend both on cognitive and behavioral abilities of the students, as we discuss further below.

4.2 Single-Dimensional Ability Example

To illustrate how identification works with the factor structure assumption, first consider the simplest example in which only one factor (e.g., the first element of θ_i: $\theta_{i,1}$) affects the outcome and selection equations in period 1, i.e., the standard case in which one assumes that unobserved ability is uni-dimensional. Suppose the outcome in period 1 is free of selection,[30] so

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

It is straightforward to show that the joint distribution of $\epsilon_i(1) = \theta_{i,1}\alpha_1(1) + \varepsilon_i(1)$ and $U_i(1) = \theta_{i,1}\rho_1(1) + \upsilon_i(1)$ is nonparametrically identified (e.g., Heckman and Smith, 1998). Further, normalizing $\rho_1(1) = 1$,[31] we can form[32]

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

With $\alpha_1(1)$ in hand, it follows from a Theorem of Kotlarski (1967)[33] that the distribution of $\theta_{i,1}$ (and of $\varepsilon_i(1)$ and $\upsilon_i(1)$) is nonparametrically identified. For example, suppose these

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

[31] 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.

[32] Notice that we implicitly assume that the distribution of θ is not symmetric. This is not a necessary assumption but it simplifies the proofs. In online Appendix E we show how to identify the model when the distribution is symmetric.

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

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

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

distributions are such that they can be characterized by their moments (see Billingsley, 1995 for conditions). Then, intuitively, identification of the distribution of $\theta_{i,1}$ follows from the fact that we can recover all its moments from $E\left(\epsilon_i^k(1) U_i(1)\right) = \alpha_1^k(1) E\left(\theta_{i,1}^{k+1}\right)$ for $k > 0$. Formally, one wants to characterize a distribution using its characteristic function and not moments, and this is precisely what the Kotlarski argument does.

Next consider the (selection corrected) second period outcomes

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

and selection equation

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

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

4.3 Multidimensional Abilities

We extend this analysis to the case in which unobserved ability (θ_i) is multidimensional beyond the correlated shocks (i.e., gaining a new element of θ_i each period). Associated with ability is a set of tests or markers that measure these components of ability imperfectly, which in our application correspond to the initial tests given to students in kindergarten before any grade repetition takes place. The existence of selection-free initial test scores is not crucial (provided we can correct for selection), but we keep it because a) it is common to many situations and b) it simplifies the exposition of the identification argument.[34]

We consider a normalization of θ_i such that true ability at the initial period consists of three independent components (A_i, B_i, C_i). In particular, assume we have access to $N_c \geq 2$ measures (or tests) of cognitive functions $\zeta_{i,j}$, and $N_b \geq 2$ measures of behavioral functions,

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

$\beta_{i,j}$, that are measured free of selection. As before, we keep all conditioning on covariates implicit to simplify notation. We write the j^{th} demeaned period 1 cognitive test as

$$\zeta_{i,j,1} = A_i \alpha_{\zeta,j,1} + C_i \pi_{\zeta,j,1} + \varepsilon_{i,\zeta,j,1}, \tag{5}$$

and the j^{th} demeaned behavioral test as

$$\beta_{i,j,1} = A_i \alpha_{\beta,j,1} + B_i \phi_{\beta,j,1} + \varepsilon_{i,\beta,j,1}. \tag{6}$$

Under this interpretation, tests are noisy measures of the components of ability. We take science, math and reading test scores as markers of cognitive ability C_i and general ability A_i (i.e., ζ) and the SRS ratings on students behavioral and social skills as our noisy measures of the behavioral ability B_i and general ability A_i (i.e. β). This is not to say that cognitive ability plays no role in behavioral aspects or vice versa but rather that whatever is common between these functions is captured by the general ability component A_i. The cognitive ability component C_i and the behavioral component B_i measure the part of ability that is used exclusively for the corresponding function. As we show below, this normalization is only required in the first period and all components of ability can affect all test scores, regardless of whether they are cognitive markers or behavioral markers, in all other periods. Other normalizations are possible, but the present normalization may also be applicable to other settings with multidimensional unobservables.[35]

Semiparametric identification follows similarly to the one factor model. We prove semiparametric identification of the model formally in online Appendix E**.** Intuitively, we now take moments across cognitive and behavioral equations to recover the period 1α-parameters and the nonparametric distribution of A. We then take cross moments within cognitive tests to recover the period 1 π-parameters and the distributions of C_i and $\varepsilon_{i,\zeta}$ and cross moment within behavioral tests to recover the period 1 ϕ-parameters as well as the nonparametric distributions of B_i and $\varepsilon_{i,\beta}$. In essence A_i represents everything that correlates behavioral and cognitive scores, B_i and C_i capture the residual correlation in behavioral and cognitive scores respectively after accounting for A_i. The role of ε_i is to capture the residual variance in scores not captured by (A_i, B_i, C_i). Notice that, by taking cross moments up to the identifiable distribution of $\upsilon_i(1)$, we now know the distribution of the unobservables determining who will be retained in period 1.

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

[35]See Cunha, Heckman and Schennach (2010) and Bonhomme and Robin (2010) for examples.

we can correct period 2 test scores using semiparametric selection correction methods like the control function approach.[36] We can then repeat the arguments above and recover the period 2 loadings and the distribution of the period 2 $\varepsilon's$ from the selection-corrected period 2 outcomes. However, since we now know the distribution of abilities in advance, we can let all three types of ability enter all equations (whether behavioral or cognitive) without having to normalize some loadings to zero. That is, the normalization that B_i only enters β-equations and C_i only enters ζ-equations need only apply to the first period.

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

The factor structure we impose has other advantages. For example, we correct for potential biases due to selective sample attrition (e.g., children moving to a different school if they know they will be retained in their current school) by adding an equation for missing data (say a binary model for attrition) that depends on the same common vector θ_i.

5 Estimation

In order to take our model to the data, we further specify our estimating equations as linear in parameters as follows. 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[37]

$$\zeta_{i,j,1} = X_{i,1}\gamma_{\zeta,j,1} + A_i\alpha_{\zeta,j,1} + C_i\pi_{\zeta,j,1} + \varepsilon_{i,\zeta,j,1} \tag{7}$$

and

$$\beta_{i,j,1} = X_{i,1}\gamma_{\beta,j,1} + A_i\alpha_{\beta,j,1} + B_i\phi_{\beta,j,1} + \varepsilon_{i,\beta,j,1}. \tag{8}$$

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

[37] We follow the identification arguments in Section 4.1 and, without loss of generality, impose the following normalizations. We normalize the general ability loading on the first period general knowledge test to 1, so A can be interpreted as a trait that is associated positively with higher scores in the general knowledge test. The loading on cognitive ability is normalized to 1 on the first period math test, so C is associated with higher math scores. Finally, we normalize the behavioral loading on the self-control marker to 1.

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

$$\zeta_{i,j,t} = X_{i,t}\gamma_{\zeta,j,t} + A_i\alpha_{\zeta,j,\infty,t} + B_i\phi_{\zeta,j,\infty,t} + C_i\pi_{\zeta,j,\infty,t} + \sum_{\tau=2}^{t}\eta_i^{(\tau)}\delta_{\zeta,j,t}^{(\tau)} + \varepsilon_{i,\zeta,j,t}$$
$$+ \sum_{r=1}^{t-1} D_i(r)\left[\Phi_{t,r} + A_i\left[\alpha_{\zeta,j,r,t} - \alpha_{\zeta,j,\infty,t}\right] + B_i\left[\phi_{\zeta,j,r,t} - \phi_{\zeta,j,\infty,t}\right] + C_i\left[\pi_{\zeta,j,r,t} - \pi_{\zeta,j,\infty,t}\right]\right]. \tag{9}$$

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

The actual form of the model for retention we use is the following.[38] We write the latent index V as

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

$D_i(R_i)$ would then be defined as

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

Notice that, consistent with our data, we allow for exclusions in the index, so that some variables (Z) are included in the retention equations but not in the outcomes. In the data this corresponds to 7 binary measures of the retention policies summarized in Table 2.[39] As discussed in Section 4.1, given that test scores in kindergarten are free of selection, the

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

[39] We examine whether these are valid exclusions in a simple two stage least squares regression and find that they satisfy the test of overidentifying restrictions in this setting. We also estimted the model with and without the exclusion restrictions and performed a loglikelihood test where we cannot reject that the exclusion restrictions are important.

additional assumption of valid exclusion restrictions is not necessary, but rather aids in identification. Similarly, given valid exclusions, the assumption of initial test scores free of selection is not necessary for identification. Furthermore, to address non-random sample attrition, we also include a similar selection equation for students who select out of the sample.

The distributions of the unobservables $(A, B, C, \{\eta^{(\tau)}\}_{\tau=2}^{\bar{t}}, \varepsilon, \upsilon)$ in the model are non-parametrically identified, as shown in Section 4.1. However, for estimation purposes, we specify all of the distributions and allow them to follow mixtures of normals with either two or three components. Furthermore, while our identification arguments are presented in a sequential fashion and lead naturally to a multi-step estimation procedure, we estimate all of the parameters in the model jointly by maximum likelihood in a single step.

6 Results

Turning to the results, we begin by estimating average treatment effects and treatment on the treated parameters for retained students in the last year of our data, 2003/04. Next, we consider how effects vary by the abilities of the students. Then, we turn to the question of time-varying treatment effects, i.e., how estimated effects vary based on the time elapsed since treatment. We put our estimates in context by contrasting our estimate of the average treatment effect at different periods to estimates from OLS and fixed effect models. Finally, we perform a policy experiment that considers the effects of a marginal change in retention policy.

To begin, we find that our model fits the means and variances of all the test measures very well. We cannot reject that the values predicted by the model equal those in the data. The same is true for the probabilities of retention in the data. Furthermore, we cannot reject the hypothesis of equality of predicted and actual probabilities.[40]

Figure 1 presents evidence of selection on the different components of ability. Ignoring kindergartners for the moment, we find that for all abilities the ordering is such that early retainees have lower ability than later retainees who have lower ability than students who are not retained. This is consistent with a dynamic selection model in which you first retain the lowest ability students and then in the next round the next lowest ability, etc. Kindergarten retention appears to be an exception in that they are higher ability than early retainees but lower ability than late retainees. This may follow because the decision to retain children in kindergarten is different than in other grades and is not as closely related to ability.

[40] In online Appendix Tables D1 and D2 we present evidence of the fit of the model. Parameter estimates and standard errors are available in online Appendix D.

This evidence provides important support for our method, i.e., the need to account for both dynamic selection and multidimensional abilities.

Table 4 describes one parameter of interest–the treatment on the treated (and the untreated) parameters for both reading and math test scores (Panels A and B respectively) in the last year in our data, the 2003/04 school year.[41] The columns correspond to actual treatment statuses, whereas the rows compare potential gains across treatment statuses relative to not being retained. In other words, the first row describes the treatment effect of being retained in kindergarten versus not being retained. The last column describes the average treatment effects.

Considering first the treatment on the treated parameters, students who are actually retained in kindergarten perform 6% lower in reading and math by 2003/04 than if they had not been retained. This does not mean that students who are retained lose acquired knowledge by being retained. It means that by age 11 (i.e., in 2003/04) a pair of identical students (one of whom was retained) would both have higher test scores than they did at age 6. The retained student's age 11 score, however, would be 6% lower than his counterpart. Students who are retained early perform about 11% lower in reading and 10% lower in math than if they had not been retained. The results for late retention vary across math and reading, with late retainees experiencing gains of 2% in reading but losses of 5% in math, although these results are not statistically significantly different from 0.

Overall, the treatment on the treated parameters suggest that the effect of retention is generally negative. In contrast, the average treatment effects reported in the last column predict that the effect of retention in kindergarten is small or 0 and positive for early retention. Again, the effect is not statistically significantly different from 0 for late retention. We can see that these non-negative average treatment effects are driven by the untreated students, for whom the treatment effect of retention is generally positive. Below we provide some intuition behind this finding.

6.1 Heterogeneity in Treatment Effects by Abilities

An advantage of our method is that we can provide direct estimates of how treatment effects vary by the unobservable abilities of the student. This also sheds light on the disparities between average treatment effects and treatment on the treated described in Table 4. Figure 2 shows how the treatment effect of being retained at different grades varies across the percentiles of the general, behavioral and cognitive ability distributions for reading and math for the 2003/04 academic year, when children are approximately age 11. Comparing

[41] The predicted levels of achievement from which these gains are calculated are included in online Appendix Table D5.

across graphs, we see that generally lower ability students experience losses (or are no better off) due to retention whereas the higher ability students benefit from retention. Thus, what the main pattern shows is (the perhaps surprising finding) that a high ability student would actually perform better by 2003/04 if retained relative to not being retained and receiving an additional grade's worth of course material. The opposite is true for low ability students.

There could be several reasons for these findings. First, it may not be possible to estimate the effect for high ability students if we do not observe high ability students being retained in our data. However, it is important to note that the test scores reported in the ECLS-K are not actually those used to determine retention decisions. While we recognize a student as high ability from the factor decomposition of the history of his performance on these tests, his performance in the classroom could suggest otherwise. Even if we restrict the sample to students whose measured achievement is below the median, this sample does not capture all retainees. Furthermore, we test that the results for high-ability students are not just noise; confidence intervals show that the effects are statistically significantly different from 0 in general.

A second potential reason is that our model is restricted to be linear in ability. It could be that in reality the students close to the margin benefit, while high and low ability students experience losses from retention. We estimate a more flexible version of our model that allows for a quadratic in ability in the outcome equations, thus permitting this sort of inverted-U-shaped pattern in ability.[42] While we do find evidence of some inverted U's, this is far from being a consistent pattern. In some cases, the upward sloping treatment effects in ability become even more pronounced. Furthermore, model selection tests favor the linear model over the quadratic ones.

Another reason could be that higher ability students actually benefit more from retention than low ability students. We find that the factor loadings are larger for the retained than for the not retained outcomes and positive in cognitive and general ability (see online Appendix Table D10). Given that ability has mean 0, this means roughly that high ability students experience achievement gains relative to not being retained, whereas low ability students experience losses relative to not being retained. There are several intuitively-appealing explanations for this that are supported in our data and in the literature.

First, high ability students may have higher ability parents (assuming intergenerational transmission of human capital). We find evidence of this in our data; higher ability students who are retained in kindergarten come from higher SES families and are less likely to be

[42]We do this in two ways. First, we estimate a model that incorporates a quadratic in each ability. Simulations support that this model is identified, although we cannot show identification analytically. Second, if we permit only the cognitive test scores to be quadratic in abilities, we can at least show that we have enough equations for the number of unknowns (i.e., the order condition).

from single-parent families. Higher SES parents may be better-equipped to ensure that when their child is retained he gets the best teachers and the attention (and resources) he needs. Thus, resources may be invested disproportionately more in high ability students who are retained than in low ability students. We test for this directly using a difference in difference strategy and find some evidence to support this hypothesis. Higher ability students who are retained in kindergarten experience larger increases in the quantity of books in the home in the next year and are more likely to have a TV rule put in place, relative to lower ability students who are retained.

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

Additionally, even if teachers and/or parents put more resources into students who are retained equally, we may still observe this pattern. If high ability students are better-equipped to take advantage of these additional resources than low ability students, this may explain the difference across ability types.

High ability students also may benefit from being retained if, by being retained, they are put in the position of teaching other students or gain confidence as they see that they are able to perform well next to the new cohort of students. In contrast, low ability students who are retained may not be in a position to offer help to their new cohort of peers. They may even lose self-esteem if they find that they continue to perform worse next to their younger cohort. This finding is supported by Bedard and Dhuey (2006) and others suggesting that the age relative to other children in the classroom matters for performance.

Importantly, we should emphasize that while we are able to provide support for our finding that high ability students who are retained benefit relatively more than lower ability students, we would not conclude from our findings that in general high ability students should be retained for several reasons. First, we we can only estimate the effect of retention on the support of students who are actually retained. While there appear to be some relatively high ability students retained in our data, as argued above, the results may not generalize to the highest ability students or the average high ability student. Second, the negative consequences of the year lost by a high ability student from retention in terms of wages and additional schooling could easily outweigh the achievement benefits we estimate in our data. Third, the model is not a general equilibrium model and clearly could not accurately predict the effect of retaining all high ability students.

6.2 Time-Varying Treatment Effects

The results so far also illustrate considerable heterogeneity in treatment effects across retention times. On the one hand, this heterogeneity would follow if there is something substantively different about retention at these different grades, such as the repetition of first grade producing larger benefits on average than the repetition of kindergarten. On the other hand, it could be that the disparities are driven by the time elapsed since retention and our choice to focus on 2003/04 outcomes. For instance, for the case of late retention, the results reported in Table 4 and Figure 2 are short run effects, achievement gains 1 to 2 years after retention. For kindergarten retention, the effects are longer run, i.e., 4 to 5 years after treatment.

To consider how treatment effects vary over time, Figures 3 and 4 compare treatment effects of kindergarten and early retention at the different periods we observe in the data. The left hand side figure depicts the evolution over time of the average treatment effect and the right hand side figure depicts the treatment on the treated for kindergarten and early retention respectively.[43] Figure 3 shows that the initial effect of being retained in kindergarten is fairly strongly negative, with students performing on average 26% lower in reading and 12% lower in math than if they had not been retained. However, 2 years later (in 2001) the average treatment effect is somewhat positive at 4%, and goes down to 3% for reading and 0 for math in 2003. Thus, while the initial effect of retention is negative and large, students on average appear to catch up in the long run.

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

With early retention, we can only compare the short run effect (in 2001) to the effect 2 years later (in 2003). In contrast to kindergarten retention, the initial effect of early retention for the average student is much smaller, approximately 0 for reading and -5% for math. The longer run effect is positive, 5% for reading and 7% for math, on average. Furthermore, the initial effect of early retention for early retainees is worse in reading than in math, -15% and -7% respectively. Notice that the initial shock for early retention is much smaller than

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

for kindergarten. This could potentially be explained by the fact that early retainees can be retained in either first or second grade, so their initial effect may be up to 2 years after retention occurred. As in kindergarten, there is evidence that reading score gains catch up over time. This does not appear to be the case for math.

The fact that the average treatment effect is, in general, less negative than the treatment on the treated over time is consistent with our findings in Section 6.1 that the effect on the treated student is more negative than for the average student.[44] Overall evidence from considering the time elapsed since treatment suggests that students begin to recover from the initial negative shock from retention 2 years later (with the exception of early retainees in math). There is also evidence that the gains may level off over time, with the treatment effects remaining negative for the treated in our sample period. Interestingly, these findings contrast to evidence in the literature which suggests that any gain in achievement from retention may actually be short-lived (Jimerson, 2001).

6.3 Comparison with estimated ATE using OLS and FE

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

Considering reading scores in panel A, the initial effect of kindergarten retention on reading in 1999/00 is negative and takes similar values across estimation methods, ranging from -24% with OLS, -26% with our method and -28% using individual fixed effects. However, by 2001/02 (column 3) the results become qualitatively different across the methods. OLS predicts that achievement is 7% lower for students retained in kindergarten, whereas our model predicts that it is 4% higher. The fixed effect estimate is approximately 0. Similarly, OLS predicts a bigger negative initial effect of early retention of -15%, in contrast to smaller estimated effects of -5% for fixed effects and 0 for our model. One reason these estimates may diverge over time is because of the changing importance of different components of ability over time (as evidenced in the variance decomposition in Tables D3 and D4). OLS and fixed effects only control for unobservable abilities in one dimension, through contemporaneous

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

test scores in the other subject for OLS and repeated values of tests in the same subject for the fixed effects. In contrast, the measure of ability in our model takes into account the whole history of test scores, as well as controlling for different dimensions of ability. The fixed effects estimator also assumes that this fixed ability component affects selection in the same way over time, which we find not to be the case using our method.

By 2003/04, OLS still estimates a negative effect of kindergarten and early retention, though the negative effect of early retention is smaller in magnitude than the initial effect in 2001/02. In contrast, the fixed effect estimator predicts a positive effect of kindergarten and early retention. Our model also predicts positive effects, but they are smaller in magnitude than the fixed effects. At the very least, this comparison suggests our findings of positive average treatment effects are not unique to our model. Even more importantly, however, OLS generally predicts the wrong sign of the average treatment effect, particularly in the long run, which would lead to the erroneous conclusion that the effect of grade retention for the average student is negative. In contrast, fixed effects overstates the benefit of grade retention for the average student in the long run, by as much as 15% higher returns than our model.

6.4 Marginal Policy Change

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

We consider the effect of a marginal change in retention policies in Table 6. In particular, we simulate the effects of changing the retention policy dummies in Table 2 to take value 0, making it harder for all schools to retain students. We present three sets of results. In column 3, we show the gains in achievement for those students who are no longer retained as a consequence of the policy change. For comparison, column 4 shows the average counterfactual gain to not being retained for students in the original retention status (i.e., the negative of the treatment on the treated parameter in Table 5), while column 5 shows the average counterfactual gain to not being retained for students who are not retained (i.e., the negative of the treatment on the untreated parameter).

For example, the first row of panels A and B, considers the case where students are originally retained in kindergarten but are now no longer retained because of the policy change for reading and math respectively. In column 3, we see that these marginal students gain 3% in both reading and math from the change in retention status to not being retained.

In contrast, the average student who is not retained would lose 3% in reading and 1% in math by not being retained relative to being retained in kindergarten. The average student already being retained in kindergarten would gain 6% in reading and in math if he were not retained. Except for the case involving late retention in reading, where the estimate is very imprecise, the point estimate of the effect for the marginal student affected by the policy lies in between the average effects for students in the original and new retention statuses.

The return to the marginal student is closer to the treatment on the treated estimate than it is to the treatment on the untreated one. This is to be expected since there is a wider range of abilities in the untreated sample. The students affected by the policy have higher abilities than the average student already retained and lower abilities than those not retained. Given the general positive relationship between ability and the benefits of retention described above, the marginal students will not benefit as much from not being retained as the average student who is already retained (i.e., the marginal students are not hurt as much by retention).

7 Conclusion

Overall, our results do not support grade retention as an effective policy for raising the performance of low achieving students. With the exception of late retainees, our estimates imply that students who are retained experience considerable achievement losses relative to not being retained, as large as 28% lower achievement than they would have acquired if they had not been retained. On the more positive side, our results suggest that retained students may catch up after several years. Yet, even if they do catch up, this would not provide strong evidence in support of retention, as the retention process is far from costless. At the very least, it may delay the student's entry into the job market by one year.

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

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

particularly policy relevant.

The disparity between the treatment on the treated and treatment effect for the average student is because of unobserved abilities. A key contribution of our approach is that it allows us to recover the distribution of the unobservables nonparametrically. Thus, we can show directly how the treatment effects vary by the abilities of the students. We find that the losses for retention are larger for low ability students. In fact, high ability students can even benefit from being retained in some cases, though as we discuss in the findings it would be a mistake to conclude from this that policy makers should retain high ability students. However, overall our results do suggest that grade retention does not improve the performance of the lowest-ability students, who are generally the targets of the policy.

Our findings also help illustrate the potential limitations of applying static methods to estimate time-varying treatment effects. Regression discontinuity designs can be a useful approach for estimating the effect of retention at a given grade. A regression discontinuity design that focuses on students close to a promotion threshold may find a positive effect of retention if the marginal students have higher ability than the average students being retained. Hence, even if lower ability students are being hurt by the policy the regression discontinuity estimate would find a positive effect. Furthermore, if there is dynamic selection, comparing these policies across grades may not be straightforward, as the students at the margin of being retained are likely to differ across grades. Interestingly, studies such as Jacob and Lefgren (2004); Greene and Winters (2007); Jacob and Lefgren (2009) which use local approaches and thus focus on marginal students present a more positive picture for grade retention than previous studies which have relied primarily on matching methods (which estimate the treatment for the average treated student rather than the marginal student).

Our findings also suggest that differences in *the* estimated effect of retention across studies (see Holmes, 1989 and Jimerson, 2001) that focus on different grades may not be surprising. One source of these disparities is simply that different types of students are retained at different grades. A second reason is that, even after controlling for dynamic selection, we find that the effect of retention varies across grades. Thus, for instance, the conclusion that first grade retention is more negative, as discussed in Alexander, Entwisle and Dauber (2003), may follow simply because early retainees are a lower ability sample and lower ability students face more negative effects of retention than higher ability students. In fact, we find that for a student who is retained early, he would not have been significantly better off by 2003/04 if he had instead been retained in kindergarten. Existing analyses are not equipped for these sorts of counterfactuals.

The method we develop extends beyond the retention application. Many policy evaluation problems involve multiple potential treatments, whether time is involved or not. These

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

References

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

Alexander, K.L., D.R. Entwisle, and S.L. Dauber. 2003. *On the success of failure: a reassessment of the effects of retention in the primary grades.* Cambridge University Press.

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

Billingsley, Patrick. 1995. *Probability and measure. A Wiley-Interscience publication.* 3 ed., New York:Wiley.

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

Carneiro, Pedro, Karsten Hansen, and James J. Heckman. 2003. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice." *International Economic Review*, 44(2): 361–422. 2001 Lawrence R. Klein Lecture.

Cattaneo, Matias D. 2010. "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability." *Journal of Econometrics*, 155(2): 138–154.

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

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

Cunha, Flavio, James J. Heckman, and Salvador Navarro. 2007. "The Identification and Economic Content of Ordered Choice Models with Stochastic Cutoffs." *International Economic Review*, 48(4): 1273 – 1309.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica*, 78(3): 883 – 931.

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

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

Greene, Jay P., and Marcus A. Winters. 2007. "Revisiting Grade Retention: An Evaluation of Florida's Test-Based Promotion Policy." *Education Finance and Policy*, 2(4): 319–340.

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

Hauser, Robert M., Carl B. Frederick, and Megan Andrew. 2007. "Grade Retention in the Age of Standards-Based Reform." In *Standards-Based Reform and the Poverty Gap.* , ed. Adam Gamoran, 120–153. Washington, DC:Brookings Institution Press.

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

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

Heckman, James J., and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data.* Vol. 10, , ed. J.J. Heckman and B. Singer, 156–245. New York:Cambridge University Press.

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

Heckman, James J., and Salvador Navarro. 2007. "Dynamic Discrete Choice and Dynamic Treatment Effects." *Journal of Econometrics*, 136(2): 341–396.

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics*, 88(3): 389–432.

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

Hu, Yingyao, and Susanne M. Schennach. 2008. "Instrumental Variable Treatment of Nonclassical Measurement Error Models." *Econometrica*, 76(1): 195–216.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics*, 86(1): 226–244.

Jacob, Brian A., and Lars Lefgren. 2009. "The Effect of Grade Retention on High School Completion." *American Economic Journal: Applied Economics*, 1(3): 33–58.

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

Jöreskog, Karl G., and Arthur S. Goldberger. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association*, 70(351): 631–639.

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

Lechner, Michael. 2004. "Sequential Matching Estimation of Dynamic Causal Models." IZA Discussion Paper 2004.

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

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

Navarro, Salvador. 2008. "Control Function." In *The New Palgrave Dictionary of Economics.*. . second ed., , ed. Steven N. Durlauf and Lawrence E. Blume. London:Palgrave Macmillan Press.

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

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

Zinth, Kyle. 2005. "Student Promotion/Retention Policies." Education Commission of the States State Notes 6551.

A Tables

Table 1: Summary Statistics

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

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

Note: For our counter-factual analyses, we only use data on students whose covariates and retention history are observable (i.e. not missing) for all time periods. Thus, we end up with fewer observations at the 2003-04 school year.

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

	Not Retained		Retained in Kindergarten		Retained Early		Retained Late	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
General Test Score	3.12	0.33	2.85	0.37	2.72	0.33	2.78	0.32
Reading Test Score	3.39	0.27	3.13	0.21	3.08	0.18	3.15	0.17
Math Test Score	3.14	0.35	2.77	0.32	2.67	0.26	2.74	0.25
Approach to Learning	0.12	0.94	-0.72	0.99	-0.91	0.95	-0.40	0.98
Self-Control	0.06	0.96	-0.31	1.02	-0.41	1.03	-0.09	0.93
Interpersonal Skills	0.06	0.96	-0.36	0.95	-0.53	1.00	-0.21	1.01
Male	0.49	0.50	0.66	0.48	0.63	0.48	0.54	0.50
Black	0.11	0.31	0.14	0.35	0.29	0.46	0.28	0.45
Hispanic	0.13	0.34	0.12	0.32	0.19	0.39	0.18	0.39
Age	5.64	0.34	5.39	0.28	5.50	0.32	5.52	0.33
Attended Full Time Kindergarten	0.57	0.49	0.62	0.49	0.61	0.49	0.72	0.45
Number of Siblings	1.39	1.08	1.65	1.27	1.80	1.41	1.52	1.25
Socioeconomic Status Index	0.13	0.77	-0.12	0.80	-0.33	0.69	-0.54	0.60
TV Rule at Home	0.89	0.31	0.90	0.30	0.83	0.37	0.90	0.31
Father in Household	0.16	0.37	0.19	0.39	0.28	0.45	0.38	0.49
Number of Books at home	82.52	60.84	71.20	60.34	50.19	49.66	45.00	42.67
Minority Students in School >25%	0.15	0.36	0.16	0.37	0.27	0.44	0.38	0.49
Public School	0.77	0.42	0.73	0.44	0.91	0.28	0.93	0.25
TT1 Funds Received by School	0.62	0.49	0.61	0.49	0.76	0.43	0.79	0.41
Teacher has a Master's Degree	0.35	0.48	0.32	0.47	0.40	0.49	0.33	0.47
Teacher Experience	14.37	9.02	14.19	9.29	13.74	8.90	12.51	9.14
Student's Class Size	20.46	4.96	19.48	5.49	20.76	4.70	20.63	4.47
Minority Students in Class >25%	0.40	0.49	0.42	0.50	0.63	0.48	0.66	0.48
Policy: Can be Retained for Immaturity	0.76	0.43	0.78	0.41	0.72	0.45	0.68	0.47
Policy: Can be Retained at Parents Request	0.75	0.43	0.76	0.43	0.79	0.41	0.76	0.43
Policy: Can be Retained due to Academic Deficiencies	0.88	0.33	0.83	0.38	0.91	0.29	0.88	0.32
Policy: Can be Retained Any Grade More than Once	0.10	0.30	0.13	0.33	0.14	0.35	0.15	0.36
Policy: Can be Retained More than Once in Elementary School	0.35	0.48	0.30	0.46	0.43	0.50	0.50	0.50
Policy: Can be Retained Without Parents Permission	0.44	0.50	0.45	0.50	0.61	0.49	0.58	0.50
Observations	7038		255		288		87	

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

Note: For our counter-factual analyses, we only use data on students whose covariates and retention history are observable (i.e. not missing) for all time periods. Thus, we end up with fewer observations at the 2003-04 school year. The last line lists the total number of usable observations (i.e. observations that contain at least one test/rating). Hence, the number of usable observations for any particular test/rating does not necessarily correspond to the number of observations in the last line. Notice that the last line does not sum to the total number of observations in table 1 (7832). This is because we don't know every childrens' retention status. Regardless, these observations can still be used in period 1, when no selection has taken place.

Table 3: Evidence for Dynamic Selection and Treatment Effect

Panel A: Reading Score

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

Panel B: Math Score

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

* Statistically significant at 5% level

[#] 1998-99 School Year

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

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

Table 4: Average Test Score Gain by Retention Status: 2003-04 School Year

Panel A: Reading Score

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

Panel B: Math Score

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

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

Table 5: Estimated Coefficients for Retention Variables in Outcome Equation

Panel A: Reading Score

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

Panel B: Math Score

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

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

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

Panel A: Reading Score

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

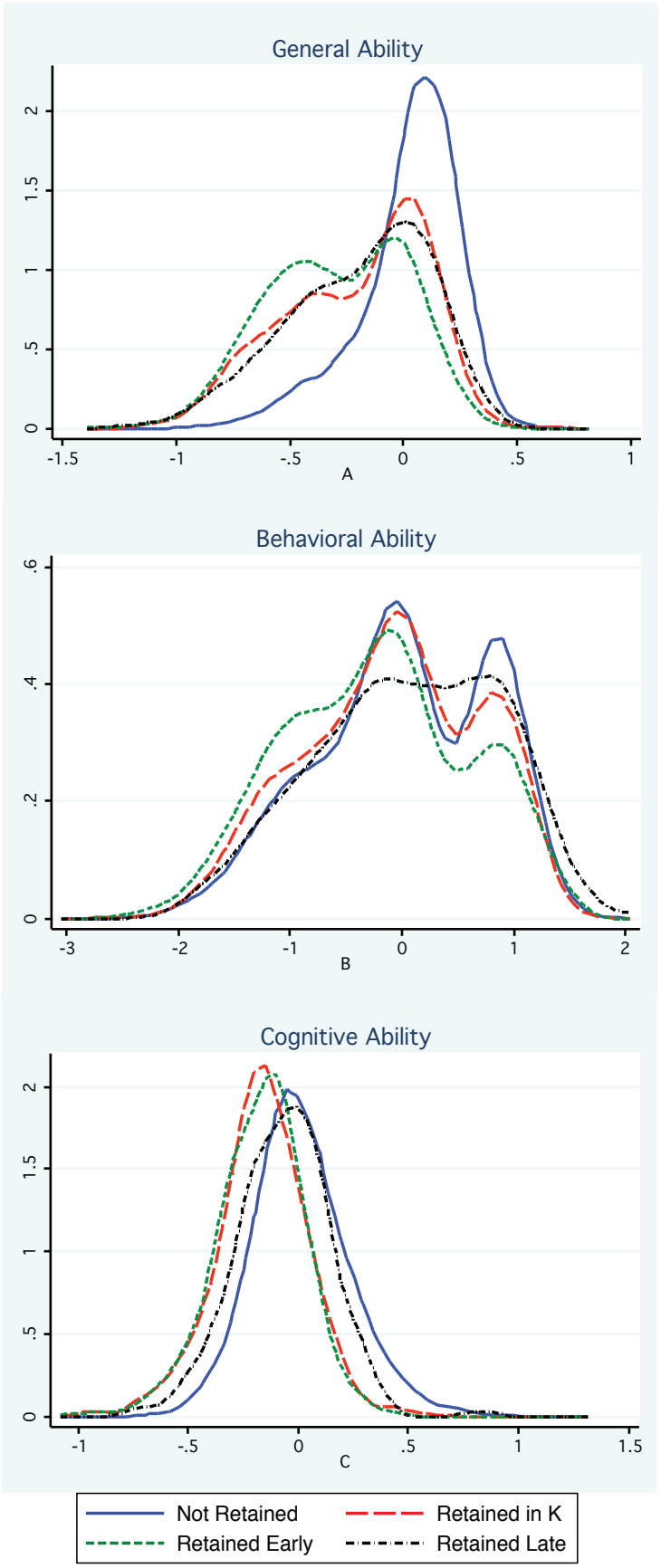
Panel B: Math Score

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

Note: We fix all retention policy variables in Table 2 to 0 for all individuals. That is we make it harder for children to be retained. Let R_0 denote the retention status under the old policy and let R_1 be the retention status under the new policy. Let ζ_0 denote the test score under original policy and ζ_1 denote the test score under the new policy. Column 3 reports $E(\zeta_1-\zeta_0 \mid R_1 \neq R_0, R_1=\infty)$, column 4 reports $E(\zeta_1-\zeta_0 \mid R_0)$ and column 5 reports $E(\zeta_1-\zeta_0 \mid R_1=\infty)$. Notice that while some people switch to other states besides $R_1=\infty$ as a consequence of the policy, there are very few and the results are harder to interpret so we focus only on the $R_1=\infty$ subgroup.

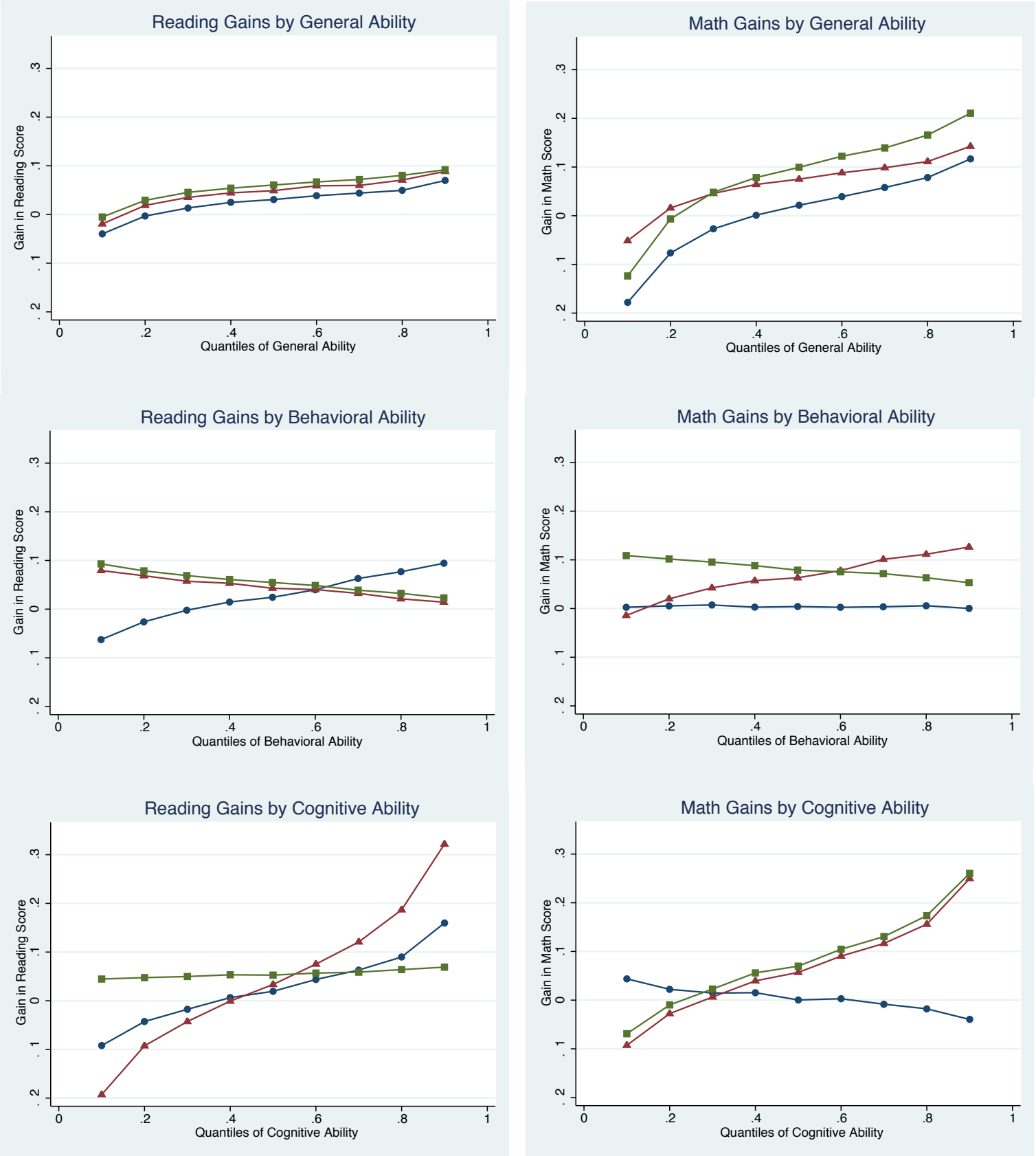
B Figures

Figure 1: Densities of Abilities by Retention Status



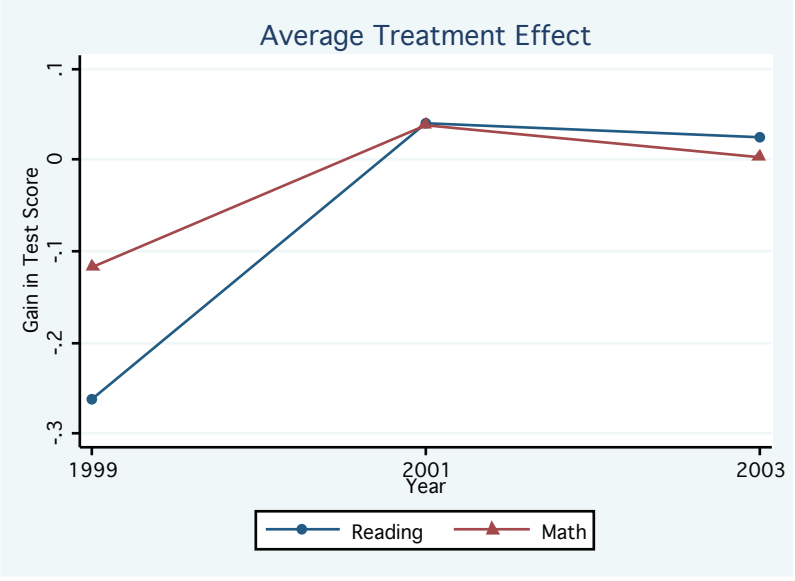
Note: Let f(X) denote the probability density function of ability X={A,B,C}. We allow f(X) to follow a mixtureof normals distribution. Let R={1,2,3,∞ } denote retention status: retained in kindergarten, retained early (1 or 2 grade), retained late (3 or 4) and not retained. The graph shows f(X|R=r) for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

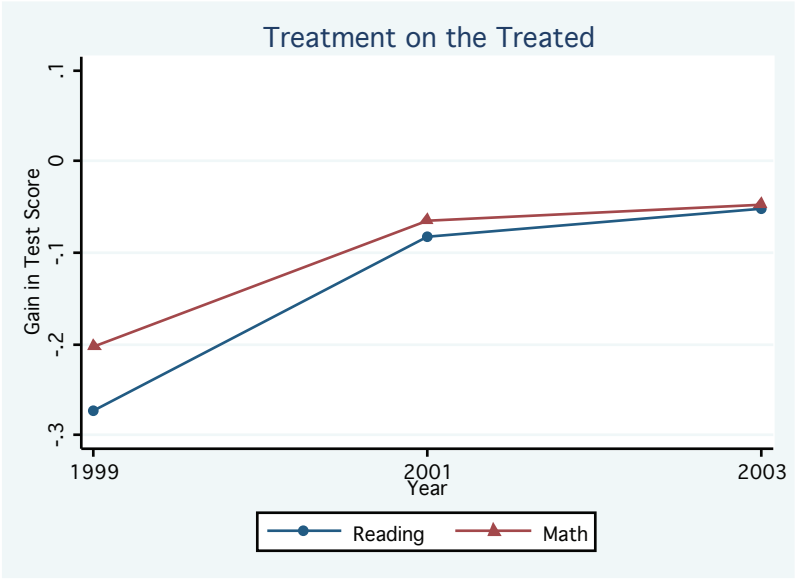
Figure 2: Achievement Gains in 2003/04 by Ability Quantiles



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

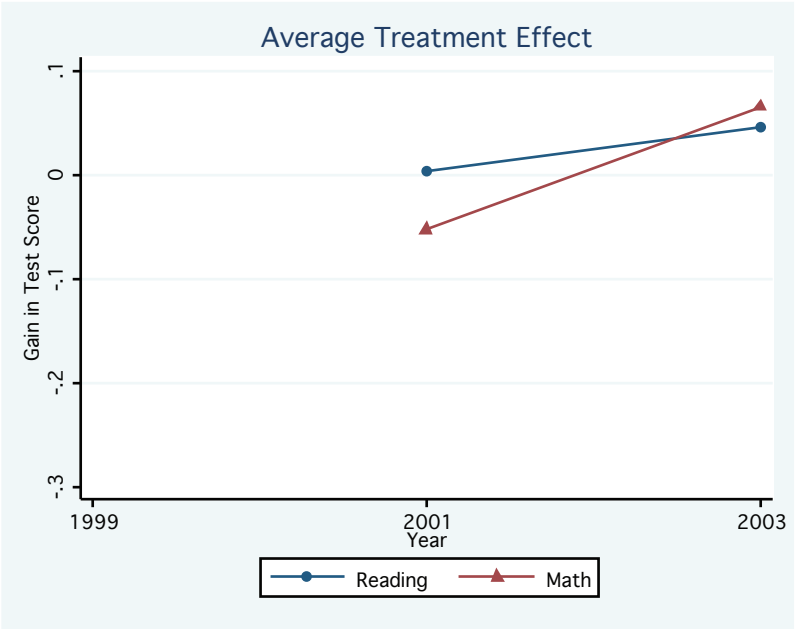
Figure 3: Achievement Gains for Kindergarten Retention over Time

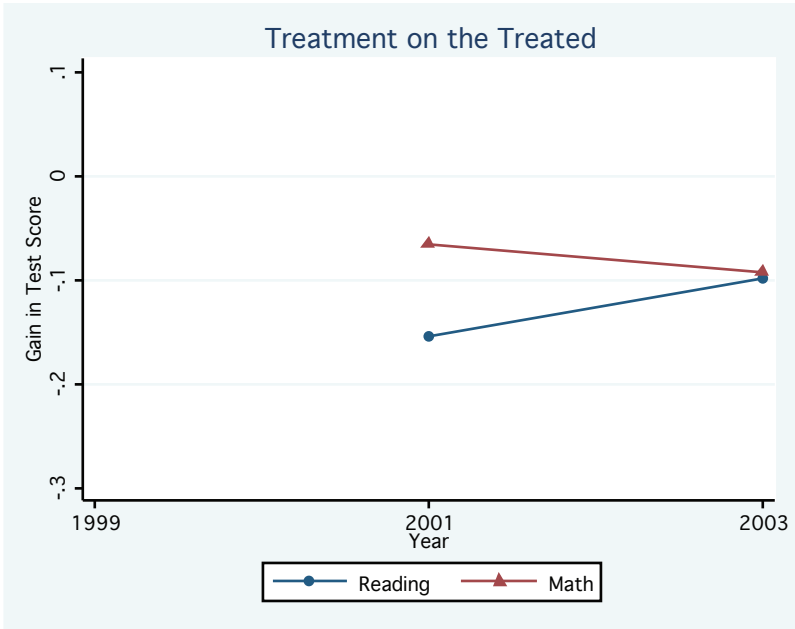




Note: Let $\zeta(t,1)$ and $\zeta(t,\infty)$ be the potential test scores at time t if the student is retained in kindergarten and if the kid is not retained at all, respectively. Let R={1,2,3} indicate the period a student is retained at. The Average Treatment Effect graph shows $E[\zeta(t,1)-\zeta(t,\infty)]$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

Figure 4: Achievement Gains for Early Retention over Time





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

C Other Methods (Online Appendix, Not For Publication)

C.1 Experimental Data

Consider designing an experiment to recover some of the different population average parameters described above.[45] 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 may have a higher potential benefit than the average 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 consider the simple three period example of Section 4. Let $S_i = \infty$ if an individual is randomized into not receiving treatment and $S_i = t$ if the individual is randomized into receiving treatment at time t. Table C1 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 $V_i(1) > 0$ or $V_i(1) < 0$). Then, we take the individuals who would under normal circumstances receive treatment $R_i = 1$ (i.e., $V_i(1) > 0$) and randomize them into receiving treatment at $t = 1$, at $t = 2$ or not receiving treatment. In terms of our example, we observe children who would be retained in kindergarten and children 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 $R_i = 1$ ($V_i(1) > 0$).

We then go on to next period and we let those individuals who were not selected into treatment at 1 ($V_i(1) < 0$) be selected into either $R_i = 2$ ($V_i(2) > 0$) or into no treatment ($V_i(2) < 0$). We then randomize them into either receiving treatment or not. We cannot

[45] 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.

randomize people into getting treatment $R_i = 1$ (elements in bold in Table C1) since that can only be done in period 1, but at period 1 we did not know whether they would be selected into $R_i = 2$ or $R_i = \infty$. 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. Depending on the goal of the analysis, these may not be a problem. Other experiments can be designed that recover versions of these parameters that are of relevant for policy design. For example, one could randomize conditional on the estimated probability of being retained in period 2 conditional on the information available at period 1. In this case a parameter measuring the effect of retaining in period 1 children who are highly likely to be retained in period 2 can be recovered.

The second lesson that this simple example delivers, one particularly important from a practical perspective, is that the data requirements of experiments are much larger in the dynamic case than in the static case. This is 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. Both experiments and methods based on conditional independence (e.g., matching-type methods) are data hungry, more so for the dynamic case.

C.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. In order to help fix ideas, we continue with the simple 3 period example of Section (4) 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.

C.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 of the unobservables. Take, for example, the second unobservable term in equation (4). In this case we have

$$E\left(D_i\left(1\right)\left[\epsilon_i\left(3,1\right)-\epsilon_i\left(3,\infty\right)\right]|Z_i\right)=E\left(\epsilon_i\left(3,1\right)-\epsilon_i\left(3,\infty\right)|Z_i,D_i\left(1\right)=1\right)\Pr\left(D_i\left(1\right)=1|Z_i\right)$$

in the unobservables. Even though we assume $E\left(\epsilon_i\left(3,T\right)|Z\right)=0$ for all T, $E(\epsilon_i\left(3,1\right)-\epsilon_i\left(3,\infty\right)|D_i\left(1\right)=1,Z_i)$ will usually not equal zero since now we are also conditioning on $D_i\left(1\right)=1$ and the decision 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 heterogeneity within the instrumental variables framework by using local methods like the Local Average Treatment Effect ($LATE$) of Imbens and Angrist (1994) and regression discontinuity designs (Hahn, Todd and Van der Klaauw, 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 and Lefgren (2004) and Nagaoka and 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 $R_i=1$ but not at $R_i=2$. Assume also that we have an instrument Z that affects the choice of whether to receive treatment at time $R_i=1$ but does not affect the outcomes. Furthermore, assume that Z can take two values, $z_2>z_1$ such that $E\left(D_i\left(1\right)|Z_i=z_2\right)>E\left(D_i\left(1\right)|Z_i=z_1\right)$ 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 $D_i(1, z_2)$ be the indicator of whether an individual gets treatment at period 1 when $Z_i = z_2$ and define $D_i(1, z_1)$ accordingly. In this binary treatment case the LATE parameter is given by

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

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 $D_i(1, z_1) = 0$ 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 non-compliers, i.e., inframarginal individuals for whom $D_i(1, z_2) = D_i(1, z_1) = 0$, drop from the LATE calculation, in the dynamic case it may be the case that $D_i(2, z_2) \neq D_i(2, z_1)$. 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.[46]

By imposing strong restrictions on the selection process (mainly that $U_i(r) = U_i$ for all r), an alternative is the local instrumental variables approach to ordered choice models of Heckman and Vytlacil (2007*b*) 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 the standard LATE approach described above at the cost of requiring a very particular type of instrument and decision process.

C.2.2 Control Function

An alternative to instrumental variables methods is to use the control function approach which models the selection process explicitly,[47] extending it to account for dynamics. Let

[46] See Angrist and Imbens (1995) for a similar result in a model with multiple treatments.
[47] See Heckman and Robb (1985) and Navarro (2008).

$P_{i,1}$ denote the probability of getting treated at $R_i = 1$. The event $R_i = 1$ can be written as

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

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

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

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 and Navarro (2007) other semiparametric restrictions are possible. Once $K_1(P_{i,1})$ is recovered one can apply the law of iterated expectations to get

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

The only unknown term in this expression is $E(\epsilon_i(3,1) | R_i \neq 1)$ so we can solve for it. However, as with the case of experimental data neither $E(\epsilon_i(3,1) | R_i = \infty)$ nor $E(\epsilon_i(3,1) | R_i = 2)$ can be recovered. Because $R_i = 2$ 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 using a similar reasoning (i.e., by forming control functions for $R_i = \infty$ and $R_i = 2$ 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 modeling 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 propose in this paper, are needed.

A common difficulty of control function methods will become more relevant in the dynamic setting. In particular, notice that in order to recover the control function a large support restriction (i.e., identification at infinity) is necessary in order to separate the con-

stant in K_1 from the one in $\Phi(3,1)$. As more periods and potential treatment are introduced more and more of these large support restrictions will need to be satisfied.

C.3 References

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

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

Heckman, James J., and Edward J. Vytlacil. 2007*b*. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments." In *Handbook of Econometrics, Volume 6.* , ed. J. Heckman and E. Leamer. Amsterdam:Elsevier.

Heckman, James J., and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data.* Vol. 10, , ed. J.J. Heckman and B. Singer, 156–245. New York:Cambridge University Press.

Heckman, James J., and Salvador Navarro. 2007. "Dynamic Discrete Choice and Dynamic Treatment Effects." *Journal of Econometrics*, 136(2): 341–396.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics*, 86(1): 226–244.

Nagaoka, Jenny, and Melissa Roderick. 2005. "Retention Under Chicago's High-Stakes Testing Program: Helpful, Harmful, or Harmless?" *Educational Evaluation and Policy Analysis*, 27(4): 309–340.

Navarro, Salvador. 2008. "Control Function." In *The New Palgrave Dictionary of Economics.*. . second ed., , ed. Steven N. Durlauf and Lawrence E. Blume. London:Palgrave Macmillan Press.

Nekipelov, Denis. 2008. "Endogenous Multi-Valued Treatment Effect Model under Monotonicity." Unpublished manuscript, Berkeley.

D Supplemental Tables (Online Appendix, Not For Publication)

Table D1: Predicted and Data Means and Standard Deviations of Kindergarten (1998-99 School Year) Test Scores/Ratings

Test Scores / Ratings	Predicted Mean	Predicted Standard Deviation	Actual Mean	Actual Standard Deviation
Reading Test Score	3.355 (0.011)	0.281 (0.002)	3.364	0.281
Math Test Score	3.096 (0.005)	0.361 (0.002)	3.103	0.362
Approach to Learning Rating	0.005 (0.013)	0.976 (0.009)	0.047	0.976
Self-Control Rating	-0.001 (0.015)	0.969 (0.008)	0.029	0.971
Interpersonal Skills Rating	-0.017 (0.015)	0.973 (0.008)	0.018	0.976

Note: Behavioral measures are standardized to have mean zero and variance equal to one. These predicted values are calculated from 50000 simulations based on the estimated model. Bootstrap standard errors are in parentheses.

Table D2: Predicted and Data Retention Probabilities (Conditional on Survival)

	Data	Model	
		Predicted	Standard Error
Retained in Kindergarten	3.364%	3.579%	0.325
Retained Early (1st or 2nd grade)	4.012%	4.245%	0.466
Retained Late (3rd or 4th grade)	1.325%	1.354%	4.802

Note: The table caclulates the probability of retention at t, conditional on not having been retained before t. Standard Errors obtained via 1000 bootstrap replications.

Table D3: Fraction of the Unobservable Variance Explained by Each Factor for Cognitive Scores by Time Period

Reading Test Score	Time Period 1	2	3	4	Math Test Score	Time Period 1	2	3	4
General Ability	0.186	0.233	0.440	0.453	General Ability	0.306	0.378	0.438	0.439
	(0.008)	(0.013)	(0.015)	(0.044)		(0.009)	(0.013)	(0.014)	(0.061)
Behavioral Ability	--	0.001	0.003	0.003	Behavioral Ability	--	0.000	0.001	0.002
		(0.001)	(0.002)	(0.016)			(0.000)	(0.001)	(0.014)
Cognitive Ability	0.497	0.343	0.158	0.112	Cognitive Ability	0.457	0.217	0.152	0.106
	(0.011)	(0.015)	(0.015)	(0.036)		(0.009)	(0.011)	(0.011)	(0.117)
Persistent Factor 1	--	0.068	0.130	0.166	Persistent Factor 1	--	0.088	0.202	0.219
		(0.008)	(0.010)	0.018			(0.008)	(0.012)	(0.036)
Persistent Factor 2	--	--	0.043	0.039	Persistent Factor 2	--	--	0.033	0.043
			(0.006)	(0.006)				(0.005)	(0.009)
Persistent Factor 3	--	--	--	0.001	Persistent Factor 3	--	--	--	0.000
				(0.001)					(0.000)

Note: Let V_c=var($C\alpha_c$) and let V_u=var($C\alpha_c$+$A\alpha_a$+$B\alpha_b$+ε) for the parameters of a given equation. Then the fraction of the variance explained by the cognitive factor, for example, is given by V_c/V_u. These values are calculated from 50000 simulations based on the estimated model. Bootstrap standard errors are in parentheses.

Table D4: Fraction of the Unobservable Variance Explained by Each Factor for Behavioral Measures by Time Period (1998/99 School Year)

	Rating for Approach to Learning	Rating for Self-Control	Rating for Interpersonal Skills
General Ability	0.199	0.055	0.088
	(0.009)	(0.004)	(0.005)
Behavioral Ability	0.412	0.674	0.672
	(0.010)	(0.009)	(0.009)

Note: Let V_c=var($C\alpha_c$) and let V_u=var($C\alpha_c$+$A\alpha_a$+$B\alpha_b$+ε) for the parameters of a given equation. Then the fraction of the variance explained by the cognitive factor, for example, is given by V_c/V_u. These values are calculated from 50000 simulations based on the estimated model. Bootstrap standard errors are in parentheses.

Table D5: Average Test Scores by Potential and Actual Retention Status: 2003-04 School Year

Panel A: Reading Score

Potential Retention Status \ Actual Retention Status	A student who is actually (i.e. conditional on retention status being:) Not Retained	Retained in Kindergarten	Retained Early	Retained Late	Unconditional
would obtain if the student was: Not Retained	4.969 (0.011)	4.818 (0.017)	4.737 (0.020)	4.768 (0.048)	4.953 (0.011)
would obtain if the student was: Retained in Kindergarten	5.003 (0.018)	4.761 (0.023)	4.651 (0.029)	4.745 (0.063)	4.980 (0.016)
would obtain if the student was: Retained Early	5.028 (0.022)	4.726 (0.028)	4.626 (0.032)	4.721 (0.080)	4.999 (0.020)
would obtain if the student was: Retained Late	5.028 (0.112)	4.844 (0.061)	4.753 (0.082)	4.790 (0.104)	5.009 (0.101)

Panel B: Math Score

Potential Retention Status \ Actual Retention Status	A student who is actually (i.e. conditional on retention status being:) Not Retained	Retained in Kindergarten	Retained Early	Retained Late	Unconditional
would obtain if the student was: Not Retained	4.760 (0.006)	4.593 (0.016)	4.494 (0.020)	4.535 (0.055)	4.742 (0.005)
would obtain if the student was: Retained in Kindergarten	4.771 (0.024)	4.536 (0.028)	4.410 (0.032)	4.463 (0.076)	4.747 (0.022)
would obtain if the student was: Retained Early	4.839 (0.022)	4.535 (0.026)	4.399 (0.031)	4.518 (0.082)	4.809 (0.019)
would obtain if the student was: Retained Late	4.858 (0.338)	4.518 (0.144)	4.382 (0.164)	4.483 (0.269)	4.825 (0.309)

Note: Let R = 1, 2, 3, or ∞ represent the actual retention status of a student: retained in kindergarten, retained early (at grade 1 or 2), retained late (at grade 3 or 4), or never retained, respectively. Let $\zeta(i)$ be the potential test score at 2003-04 school year if the student were retained at time i=1,2,3,∞. The row i, column j element of this table calculates $E[\zeta(i) \mid R=j]$. For example, a student who was actually not retained would get 4.839 in math score on average if the student were retained at 1 or 2 grade instead. Values in parentheses are bootstrap standard errors.

Table D6: Average Test Scores by Potential and Actual Retention Status: 2001-02 School Year

Panel A: Reading Score

Potential Retention Status \ Actual Retention Status		A student who is actually (i.e. conditional on retention status being:)			Unconditional
		Not Retained	Retained in Kindergarten	Retained Early	
would obtain if the student was	Not Retained	4.803 (0.011)	4.594 (0.020)	4.502 (0.023)	4.785 (0.011)
	Retained in Kindergarten	4.858 (0.028)	4.504 (0.030)	4.384 (0.039)	4.829 (0.026)
	Retained Early	4.819 (0.018)	4.462 (0.033)	4.348 (0.040)	4.790 (0.016)

Panel B: Math Score

Potential Retention Status \ Actual Retention Status		A student who is actually (i.e. conditional on retention status being:)			Unconditional
		Not Retained	Retained in Kindergarten	Retained Early	
would obtain if the student was	Not Retained	4.551 (0.006)	4.305 (0.019)	4.196 (0.025)	4.530 (0.005)
	Retained in Kindergarten	4.601 (0.027)	4.235 (0.033)	4.103 (0.043)	4.571 (0.025)
	Retained Early	4.499 (0.032)	4.250 (0.024)	4.130 (0.030)	4.478 (0.030)

Note: Let R = 1,2,∞ or represent the actual retention status of a student: retained in kindergarten, retained early (at grade 1 or 2), or never retained, respectively. Let $\zeta(i)$ be the potential test score at 2001-02 school year if the student were retained at time i=1,2,∞. The row i, column j element of this table calculates $E[\zeta(i) \mid R=j]$. For example, a student who was actually not retained would get 4.499 in math score on average if the student were retained at 1 or 2 grade instead. Bootstrap standard errors are in parentheses.

Table D7: Average Test Scores by Potential and Actual Retention Status: 1999-2000 School Year

Panel A: Reading Score

Potential Retention Status \ Actual Retention Status		A student who is actually (i.e. conditional on retention status being:) Not Retained	A student who is actually (i.e. conditional on retention status being:) Retained in Kindergarten	Unconditional
would obtain if the student was	Not Retained	4.270 (0.011)	3.958 (0.057)	4.263 (0.011)
	Retained in Kindergarten	4.008 (0.032)	3.680 (0.065)	4.000 (0.032)

Panel B: Math Score

Potential Retention Status \ Actual Retention Status		A student who is actually (i.e. conditional on retention status being:) Not Retained	A student who is actually (i.e. conditional on retention status being:) Retained in Kindergarten	Unconditional
would obtain if the student was	Not Retained	4.055 (0.005)	3.713 (0.060)	4.047 (0.004)
	Retained in Kindergarten	3.941 (0.027)	3.520 (0.079)	3.930 (0.027)

Note: Let R = 1 or ∞ represent the actual retention status of a student: retained in kindergarten or never retained. Let $\zeta(i)$ be the potential test score at 1999-2000 school year if the student were retained at time i=1, ∞. The row i, column j element of this table calculates $E[\zeta(i) \mid R=j]$. For example, a student who was actually not retained would get 3.941 in math score on average if the student were retained in kindergarten instead. Bootstrap standard errors are in parentheses.

Table D8: Average Test Score Gain by Retention Status: 2001-02 School Year

Panel A: Reading Score

Average Gain	A student who is actually (i.e. conditional on the retention status being:)			ATE (unconditional)
	Not Retained	Retained in Kindergarten	Retained Early	
Retained in Kindergarten vs Not Retained	0.055 (0.027)	-0.090 (0.020)	-0.119 (0.027)	0.041 (0.025)
Retained Early vs Not Retained	0.016 (0.015)	-0.132 (0.023)	-0.154 (0.028)	0.004 (0.012)

Panel B: Math Score

Average Gain	A student who is actually (i.e. conditional on the retention status being:)			ATE (unconditional)
	Not Retained	Retained in Kindergarten	Retained Early	
Retained in Kindergarten vs Not Retained	0.050 (0.027)	-0.070 (0.022)	-0.093 (0.029)	0.039 (0.024)
Retained Early vs Not Retained	-0.052 (0.032)	-0.055 (0.017)	-0.066 (0.017)	-0.053 (0.030)

Note: Let R = 1, 2, or ∞ represent the actual retention status of a student: retained in kindergarten, retained early (at grade 1 or 2), or never retained, respectively. Let $\zeta(i)$ be the potential test score if the student were retained at time $i=1,2,\infty$. The row i, column j element of this table calculates $E[\zeta(i) - \zeta(\infty) \mid R=j]$. For example, the test math score of a student who was actually not retained would decrease by 0.052 if he were retained at 1 or 2 grade instead. Bootstrap standard errors are in parentheses.

Table D9: Average Test Score Gain by Retention Status: 1999-2000 School Year

Panel A: Reading Score

Average Gain	A student who is actually (i.e. conditional on the retention status being:)		ATE (unconditional)
	Not Retained	Retained in Kindergarten	
Retained in Kindergarten vs Not Retained	-0.263 (0.031)	-0.279 (0.029)	-0.263 (0.030)

Panel B: Math Score

Average Gain	A student who is actually (i.e. conditional on the retention status being:)		ATE (unconditional)
	Not Retained	Retained in Kindergarten	
Retained in Kindergarten vs Not Retained	-0.114 (0.027)	-0.193 (0.028)	-0.117 (0.027)

Note: Let R = 1 or ∞ represent the actual retention status of a student: retained in kindergarten or never retained. Let $\zeta(i)$ be the potential test score if the student were retained at time $i=1,\infty$. The column j element of this table calculates $E[\zeta(i) - \zeta(\infty) \mid R=j]$. For example, the test score of a student who was actually not retained would decrease by 0.114 if he were retained in kindergarten. Bootstrap standard errors are in parentheses.

Table D10: Estimated Factor Loadings in Cognitive Score Equations for 2003-04 School Year by Retention Status

Reading Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *	Persistent Factor 3 *
Not Retained	0.429 (0.013)	-0.002 (0.001)	0.149 (0.010)	0.098 (0.004)	0.047 (0.003)	0.007 (0.004)
Retained in Kindergarten	0.558 (0.046)	0.058 (0.016)	0.465 (0.076)	0.098 (0.004)	0.047 (0.003)	0.007 (0.004)
Retained Early	0.555 (0.043)	-0.031 (0.010)	0.805 (0.051)	0.098 (0.004)	0.047 (0.003)	0.007 (0.004)
Retained Late	0.543 (0.356)	-0.029 (0.084)	0.184 (0.500)	0.098 (0.004)	0.047 (0.003)	0.007 (0.004)

Math Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *	Persistent Factor 3 *
Not Retained	0.455 (0.014)	-0.007 (0.002)	0.243 (0.013)	0.133 (0.004)	-0.059 (0.004)	0.005 (0.002)
Retained in Kindergarten	0.800 (0.060)	-0.007 (0.009)	0.144 (0.065)	0.133 (0.004)	-0.059 (0.004)	0.005 (0.002)
Retained Early	0.681 (0.046)	0.046 (0.011)	0.677 (0.061)	0.133 (0.004)	-0.059 (0.004)	0.005 (0.002)
Retained Late	0.846 (0.479)	-0.030 (0.105)	0.670 (1.299)	0.133 (0.004)	-0.059 (0.004)	0.005 (0.002)

Science Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *	Persistent Factor 3 *
Not Retained	0.791 (0.018)	-0.028 (0.004)	0.055 (0.013)	0.103 (0.005)	0.035 (0.005)	0.005 (0.003)
Retained in Kindergarten	1.070 (0.089)	0.058 (0.023)	-0.301 (0.135)	0.103 (0.005)	0.035 (0.005)	0.005 (0.003)
Retained Early	0.800 (0.035)	0.011 (0.012)	0.130 (0.075)	0.103 (0.005)	0.035 (0.005)	0.005 (0.003)
Retained Late	0.846 (0.355)	-0.078 (0.187)	-0.145 (1.003)	0.103 (0.005)	0.035 (0.005)	0.005 (0.003)

* Loading of persistent factors are assumed to stay constant over time. Bootstrap standard errors are in parentheses.

Table D11: Estimated Factor Loadings for 2001-02 School Year

Reading Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *
Not Retained	0.542 (0.014)	-0.003 (0.001)	0.271 (0.012)	0.110 (0.004)	0.063 (0.004)
Retained in Kindergarten	0.663 (0.070)	0.071 (0.025)	0.867 (0.114)	0.110 (0.004)	0.063 (0.004)
Retained Early	0.687 (0.077)	0.009 (0.011)	0.878 (0.111)	0.110 (0.004)	0.063 (0.004)
Math Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *
Not Retained	0.589 (0.015)	-0.002 (0.001)	0.379 (0.014)	0.149 (0.004)	-0.061 (0.004)
Retained in Kindergarten	0.849 (0.093)	-0.029 (0.020)	0.702 (0.126)	0.149 (0.004)	-0.061 (0.004)
Retained Early	0.640 (0.079)	0.023 (0.013)	0.328 (0.074)	0.149 (0.004)	-0.061 (0.004)
Science Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *
Not Retained	1.032 (0.022)	-0.029 (0.005)	0.122 (0.017)	0.091 (0.006)	0.046 (0.006)
Retained in Kindergarten	1.052 (0.114)	0.000 (0.002)	0.041 (0.092)	0.091 (0.006)	0.046 (0.006)
Retained Early	0.887 0.088	0.019 0.023	0.173 0.129	0.091 0.006	0.046 0.006
Choice Equation	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *
	-2.231 (1.072)	0.074 (0.250)	-1.787 (1.526)	-0.409 (0.444)	0.075 (0.272)
Missing Equation	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *	Persistent Factor 2 *
	-0.186 (0.057)	-0.041 (0.014)	0.043 (0.025)	-0.055 (0.028)	0.054 (0.043)

* Loading of persistent factors are assumed to stay constant over time. Bootstrap standard errors are in parentheses.

Table D12: Estimated Factor Loadings for 1999-2000 School Year

General Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *
Not Retained	0.685 (0.013)	-0.012 (0.002)	0.065 (0.010)	0.022 (0.004)
Retained in Kindergarten	0.832 (0.040)	0.005 (0.011)	-0.175 (0.096)	0.022 (0.004)

Reading Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *
Not Retained	0.524 (0.016)	0.010 (0.003)	0.716 (0.021)	0.102 (0.006)
Retained in Kindergarten	0.701 (0.107)	0.042 (0.026)	0.553 (0.164)	0.102 (0.006)

Math Test Score	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *
Not Retained	0.664 (0.016)	0.001 (0.001)	0.568 (0.018)	0.117 (0.006)
Retained in Kindergarten	1.009 (0.066)	0.007 (0.018)	0.653 (0.137)	0.117 (0.006)

Choice Equation	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *
	-2.680 (0.325)	-0.282 (0.087)	-3.070 (0.602)	-0.290 (0.126)

Missing Equation	General Ability	Behavioral Ability	Cognitive Ability	Persistent Factor 1 *
	-0.186 (0.057)	-0.041 (0.014)	0.043 (0.025)	-0.055 (0.028)

* Loading of persistent factors are assumed to stay constant over time. Bootstrap standard errors are in parentheses.

Table D13: Estimated Factor Loadings for 1998-99 School Year

	General Ability	Behavioral Ability	Cognitive Ability
General Test Score	1.000[+] --	--	0.203 (0.015)
Reading Test Score	0.459 (0.013)	--	0.823 (0.014)
Math Test Score	0.745 (0.016)	--	1.000[+] --
Rating for Approach to Learning	1.662 (0.045)	0.772 (0.014)	--
Rating for Self-Control	0.888 (0.036)	1.000[+] --	--
Rating for Interpersonal Skills	1.128 (0.038)	1.002 (0.015)	--
Choice Equation	-1.580 (0.199)	-0.057 (0.049)	-1.997 (0.337)
Missing Equation	-0.186 (0.057)	-0.041 (0.014)	0.043 (0.025)

[+] Normalized to one.
Bootstrap standard errors are in parentheses.

Table D14: Estimated Coefficients for Dummy Variables in Outcome Equations

Panel A: Reading Test Score

	Outcome Equation in 1999-2000 School Year	Outcome Equation in 2001-02 School Year	Outcome Equation in 2003-04 School Year
Retained in Kindergarten	-0.263 (0.030)	0.041 (0.025)	0.025 (0.012)
Retained Early		0.004 (0.012)	0.046 (0.017)
Retained Late			0.056 (0.101)

Panel B: Math Test Score

	Outcome Equation in 1999-2000 School Year	Outcome Equation in 2001-02 School Year	Outcome Equation in 2003-04 School Year
Retained in Kindergarten	-0.117 (0.027)	0.039 (0.024)	0.004 (0.022)
Retained Early		-0.053 (0.030)	0.066 (0.019)
Retained Late			0.083 (0.309)

Panel C: General Test Score

	Outcome Equation in 1999-2000 School Year	Outcome Equation in 2001-02 School Year	Outcome Equation in 2003-04 School Year
Retained in Kindergarten	-0.019 (0.017)	--	--
Retained Early		--	--
Retained Late			--

Panel D: Science Test Score

	Outcome Equation in 1999-2000 School Year	Outcome Equation in 2001-02 School Year	Outcome Equation in 2003-04 School Year
Retained in Kindergarten	--	0.024 (0.014)	0.014 (0.027)
Retained Early		-0.043 (0.038)	0.001 (0.003)
Retained Late			0.006 (0.030)

Note: Bootstrap standard errors are in parentheses.

Table D15: Parameter Estimates for Outcome Equations

	General Test Score		Reading Test Score		Math Test Score		Science Test Score	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Constant (1998-1999)	-0.120	0.009	2.352	0.027	1.043	0.035	--	--
Constant (1999-2000)	0.049	0.007	3.075	0.031	1.658	0.039	--	--
Constant (2001-2002)	--	--	3.407	0.036	1.828	0.042	1.260	0.115
Constant (2003-2004)	--	--	3.465	0.039	1.893	0.044	1.331	0.120
Male	0.012	0.001	-0.043	0.002	0.012	0.001	0.040	0.003
White	0.057	0.004	0.000	0.001	0.009	0.001	0.014	0.002
Black	-0.061	0.006	-0.063	0.004	-0.103	0.005	-0.145	0.009
Hispanic	-0.036	0.005	-0.038	0.004	-0.054	0.004	-0.074	0.007
Body Mass Index	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Age	0.829	0.006	0.239	0.006	0.467	0.009	0.418	0.022
Age_squared	-0.050	0.001	-0.009	0.000	-0.019	0.001	-0.016	0.001
Number of Siblings	-0.016	0.001	-0.015	0.001	-0.003	0.001	-0.018	0.002
Socioeconomic Status Index	0.058	0.003	0.051	0.001	0.062	0.002	0.075	0.003
TV Rule at Home	0.022	0.004	0.011	0.003	-0.010	0.001	0.007	0.001
Mother in Household	-0.023	0.013	0.002	0.011	-0.010	0.007	-0.055	0.017
Father in Household	-0.014	0.003	-0.014	0.003	-0.007	0.002	-0.015	0.003
Number of Books at home	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Number of Books at home (squared)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Attended Full Time Kindergarten	-0.006	0.004	0.002	0.008	0.005	0.002	0.014	0.002
Minority Students in School between (1%,5%)	0.006	0.003	0.012	0.003	0.008	0.007	0.009	0.002
Minority Students in School between (5%,10%)	-0.001	0.004	0.008	0.003	0.010	0.014	0.003	0.004
Minority Students in School between (10%,25%)	-0.019	0.006	-0.010	0.001	-0.019	0.005	0.007	0.009
Minority Students in School >25%	-0.039	0.007	-0.017	0.005	-0.009	0.002	-0.049	0.013
Public School	-0.031	0.003	-0.032	0.002	-0.031	0.002	-0.038	0.005
TT1 Funds Received by School	-0.010	0.004	-0.018	0.001	-0.018	0.001	-0.007	0.004
Crime a Problem	0.005	0.002	-0.003	0.000	0.000	0.003	-0.022	0.003
Students Bring Weapons	-0.006	0.003	-0.012	0.003	-0.008	0.003	-0.012	0.003
Children or Teachers Physically Attacked	-0.005	0.004	-0.007	0.001	-0.002	0.000	0.012	0.003
Security Measures in School	-0.009	0.001	0.001	0.001	-0.007	0.001	0.006	0.005
Parents Involved in School Activities	0.009	0.001	0.002	0.003	0.002	0.000	0.006	0.001
Teacher has a Master's Degree	0.002	0.000	-0.001	0.002	0.001	0.000	0.001	0.007
Teacher Experience	0.001	0.000	0.002	0.000	0.001	0.000	0.001	0.000
Teacher Experience squared	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Student's Class Size	0.002	0.000	0.002	0.000	0.008	0.000	0.003	0.000
Student's Class Size (squared)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Teacher's Rating of Class Behavior	-0.001	0.001	-0.006	0.001	-0.005	0.001	-0.004	0.001
Minority Students in Class between (1%,5%)	0.001	0.005	0.004	0.001	-0.001	0.006	0.000	0.005
Minority Students in Class between (5%,10%)	0.009	0.007	0.003	0.003	0.010	0.003	0.015	0.003
Minority Students in Class between (10%,25%)	0.013	0.007	0.003	0.002	0.012	0.002	0.001	0.009
Minority Students in Class >25%	-0.026	0.005	0.012	0.003	-0.010	0.002	-0.020	0.010

Note: All the coefficients except for the constant term are assumed to be invariant over time.

Table D16: Parameter Estimates for Behavioral Rating Equations of the 1998/99 School Year

	Approach to Learning Rating		Self-Control Rating		Interpersonal Skills Rating	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Constant	-6.745	1.093	-2.460	0.573	-5.025	0.77
Male	-0.362	0.017	-0.262	0.016	-0.279	0.02
White	-0.047	0.023	-0.061	0.022	0.088	0.02
Black	-0.150	0.037	-0.142	0.028	-0.021	0.01
Hispanic	-0.125	0.043	-0.025	0.016	0.040	0.03
Body Mass Index	-0.007	0.002	-0.009	0.003	-0.008	0.00
Age	2.106	0.382	0.844	0.196	1.664	0.27
Age_squared	-0.148	0.033	-0.066	0.017	-0.135	0.02
Number of Siblings	-0.015	0.006	0.025	0.007	-0.014	0.01
Socioeconomic Status Index	0.112	0.012	0.036	0.010	0.063	0.01
TV Rule at Home	-0.035	0.018	-0.049	0.018	-0.005	0.01
Mother in Household	-0.248	0.070	-0.102	0.071	-0.049	0.06
Father in Household	-0.082	0.024	-0.072	0.022	-0.062	0.02
Minority Students in School between (1%,5%)	-0.028	0.022	-0.035	0.020	-0.028	0.02
Minority Students in School between (5%,10%)	0.048	0.025	-0.006	0.011	0.018	0.02
Minority Students in School between (10%,25%)	0.136	0.034	-0.012	0.021	0.078	0.03
Minority Students in School >25%	0.073	0.031	-0.135	0.027	-0.032	0.03
Public School	0.106	0.022	0.100	0.021	0.108	0.02
TT1 Funds Received by School	0.082	0.018	0.039	0.015	0.045	0.02
Crime a Problem	0.010	0.010	0.014	0.011	-0.011	0.01
Students Bring Weapons	0.015	0.017	-0.003	0.006	-0.013	0.02
Children or Teachers Physically Attacked	-0.060	0.017	-0.037	0.016	-0.094	0.02
Security Measures in School	0.013	0.011	0.019	0.012	-0.011	0.01
Parents Involved in School Activities	0.021	0.008	0.012	0.006	0.000	0.00
Attended Full Time Kindergarten	-0.058	0.016	-0.042	0.015	0.003	0.00
Number of Books at home	0.001	0.000	0.000	0.000	0.002	0.00
Number of Books at home (squared)	-0.001	0.000	0.000	0.000	-0.001	0.00
Teacher has a Master's Degree	0.012	0.011	0.049	0.016	0.051	0.02
Teacher Experience	-0.001	0.001	0.005	0.002	-0.005	0.00
Teacher Experience	0.000	0.000	0.000	0.000	0.000	0.00
Number of Kids in Class	-0.002	0.001	0.018	0.004	0.017	0.00
Number of Kids in Class (squared)	0.000	0.000	0.000	0.000	0.000	0.00
Teacher's Rating of Class Behavior	-0.091	0.010	-0.116	0.010	-0.096	0.01
Minority Students in Class between (1%,5%)	-0.018	0.026	-0.031	0.027	0.028	0.03
Minority Students in Class between (5%,10%)	-0.020	0.021	-0.004	0.009	0.037	0.02
Minority Students in Class between (10%,25%)	-0.032	0.022	-0.005	0.010	0.007	0.01
Minority Students in Class >25%	-0.006	0.008	0.020	0.016	0.114	0.02

Table D17: Parameter Estimates for Choice Equations and Missing Equations

	Choice Equation (1998-1999)		Choice Equation (1999-2000)		Choice Equation (2001-2002)		Missing Equation	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Constant (1998-1999)	6.531	1.011					-0.863	0.052
Constant (1999-2000)			-7.380	4.013			-0.879	0.064
Constant (2001-2002)					17.787	11.034	-1.230	0.090
Male	0.384	0.097	0.266	0.125	0.427	0.552	0.020	0.014
White	0.020	0.027	-0.044	0.091	0.947	0.903	-0.070	0.030
Black	-0.023	0.051	0.649	0.216	0.637	0.923	-0.021	0.028
Hispanic	-0.198	0.148	0.147	0.165	0.253	1.372	0.052	0.038
Body Mass Index	-0.034	0.016	-0.010	0.012	-0.009	0.020	0.002	0.001
Age	-1.331	0.180	2.137	1.142	-3.433	2.341	0.011	0.006
Age_squared	-0.007	0.007	-0.211	0.084	0.115	0.134	0.002	0.001
Number of Siblings	0.083	0.035	0.145	0.047	0.060	0.204	0.010	0.006
Socioeconomic Status Index	-0.189	0.070	-0.227	0.110	-0.459	0.441	0.019	0.009
TV Rule at Home	0.114	0.101	-0.300	0.186	-0.111	0.423	0.018	0.013
Mother in Household	0.125	0.291	0.243	0.407	0.148	1.202	0.018	0.062
Father in Household	0.050	0.096	-0.015	0.043	0.282	0.506	0.115	0.030
Number of Books at home	-0.002	0.002	-0.001	0.001	0.000	0.001	0.000	0.000
Number of Books at home (squared)	0.001	0.001	0.000	0.000	-0.001	0.002	0.000	0.000
Attended Full Time Kindergarten	0.234	0.097	-0.030	0.069	0.282	0.422	0.081	0.022
Minority Students in School between (1%,5%)	-0.314	0.154	-0.246	0.203	-0.303	1.123	0.115	0.033
Minority Students in School between (5%,10%)	-0.444	0.188	-0.454	0.260	0.004	0.484	0.150	0.042
Minority Students in School between (10%,25%)	-0.275	0.172	-0.363	0.346	0.025	0.902	0.302	0.054
Minority Students in School >25%	-0.453	0.194	-0.898	0.389	0.268	0.936	0.188	0.056
Public School	-0.312	0.118	0.250	0.210	0.256	0.719	0.033	0.018
TT1 Funds Received by School	-0.229	0.101	-0.280	0.160	-0.212	0.555	-0.022	0.005
Crime a Problem	0.122	0.080	0.175	0.133	-0.286	0.555	0.071	0.021
Students Bring Weapons	-0.207	0.129	0.118	0.155	0.872	0.586	0.204	0.033
Children or Teachers Physically Attacked	0.104	0.094	-0.064	0.110	0.314	0.581	0.007	0.007
Security Measures in School	0.153	0.095	-0.110	0.117	-0.075	0.387	0.005	0.005
Parents Involved in School Activities	-0.005	0.009	-0.066	0.061	-0.063	0.188	0.006	0.004
Teacher has a Master's Degree	-0.113	0.094	-0.048	0.099	-0.493	0.603	-0.001	0.002
Teacher Experience	0.002	0.002	0.000	0.003	-0.019	0.061	0.005	0.002
Teacher Experience squared	0.000	0.000	0.000	0.000	0.001	0.002	0.000	0.000
Student's Class Size	-0.068	0.024	-0.008	0.015	-0.017	0.045	0.001	0.000
Student's Class Size (squared)	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Teacher's Rating of Class Behavior	0.060	0.043	0.116	0.063	-0.011	0.081	-0.020	0.009
Minority Students in Class between (1%,5%)	0.416	0.181	-0.055	0.174	-0.963	1.236	0.097	0.036
Minority Students in Class between (5%,10%)	0.258	0.158	0.300	0.252	-0.160	0.860	0.077	0.035
Minority Students in Class between (10%,25%)	0.402	0.165	0.424	0.369	-0.520	0.891	0.173	0.042
Minority Students in Class	0.382	0.197	0.419	0.421	-0.085	0.701	0.193	0.049
Policy: Can be Retained for Immaturity	0.205	0.119	-0.320	0.165	0.437	0.701	--	--
Policy: Can be Retained at Parents Request	0.202	0.120	0.228	0.146	0.192	0.503	--	--
Policy: Can be Retained due to Academic Deficiencies	-0.490	0.145	0.363	0.304	-0.398	1.107	--	--
Policy: Can be Retained Any Grade More than Once	0.260	0.157	-0.122	0.164	0.022	0.360	--	--
Policy: Can be Retained More than Once in Elementary	-0.178	0.119	0.211	0.150	0.313	0.526	--	--
Policy: Can be Retained Without Parents Permission	0.115	0.094	0.066	0.098	-0.002	0.165	--	--
Policy: Can be Retained at any Grade	--	--	0.306	0.192	0.748	1.041	--	--

Note: All the coefficients except for the constant term are assumed to be invariant over time in the missing equation.

Table D18: Parameter Estimates of Variances of Uniqueness in Cognitive and Behavioral Outcome Equations

	Time Period			
	1998-99 School Year	1999-2000 School Year	2001-02 School Year	2003-04 School Year
General Test Score	0.0348 (0.0008)	0.0118 (0.0003)	--	--
Reading Test Score	0.0223 (0.0004)	0.0283 (0.0007)	0.0081 (0.0003)	0.0048 (0.0002)
Math Test Score	0.0266 (0.0006)	0.0247 (0.0006)	0.0086 (0.0004)	0.0060 (0.0003)
Science Test Score	--	--	0.0190 (0.0007)	0.0120 (0.0004)
Rating for Approach to Learning	0.3389 (0.0066)	--	--	--
Rating for Self-Control	0.2463 (0.0060)	--	--	--
Rating for Interpersonal Skills	0.2190 (0.0053)	--	--	--

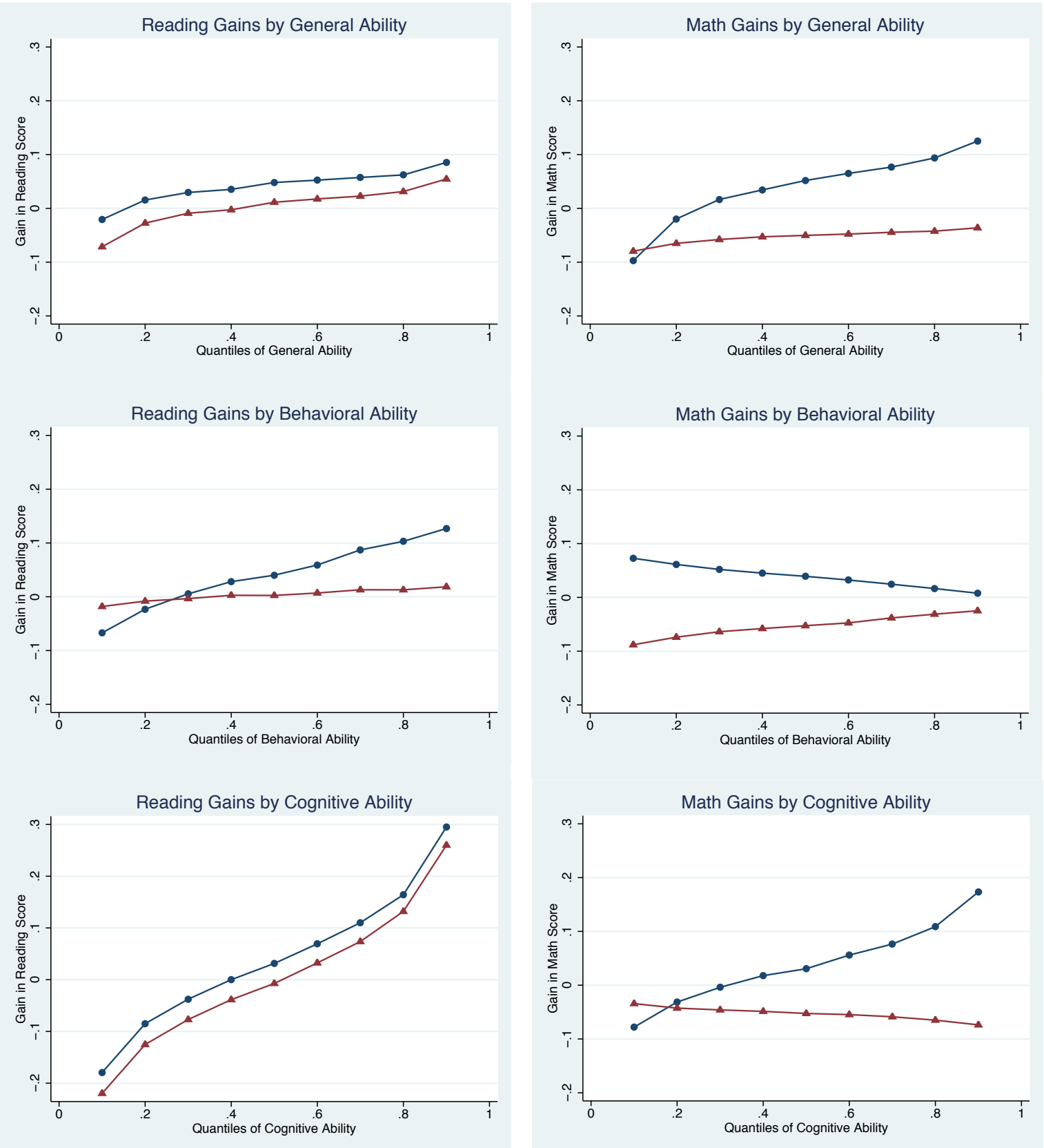
Note: Bootstrap standard errors are in parentheses.

Table D19: Parameter Estimates for Distribution of Abilities

	General		Behavioral		Cognitive	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Mean of the 1st Mixture Component	0.132	0.003	0.862	0.016	-0.009	0.001
Mean of the 2nd Mixture Component	-0.216	0.003	-0.797	0.017	-0.061	0.003
Mean of the 3rd Mixture Component	0.074	0.003	-0.006	0.017	0.069	0.003
Variance of the 1st Mixture Component	0.014	0.001	0.069	0.007	0.044	0.002
Variance of the 2nd Mixture Component	0.085	0.003	0.272	0.012	0.022	0.001
Variance of the 3rd Mixture Component	0.023	0.001	0.062	0.007	0.083	0.003
Weight of the 1st Mixture Component	0.340	0.001	0.331	0.006	0.332	0.015
Weight of the 2nd Mixture Component	0.323	0.002	0.355	0.010	0.331	0.009
Weight of the 3rd Mixture Component	0.337	0.001	0.314	0.007	0.337	0.006

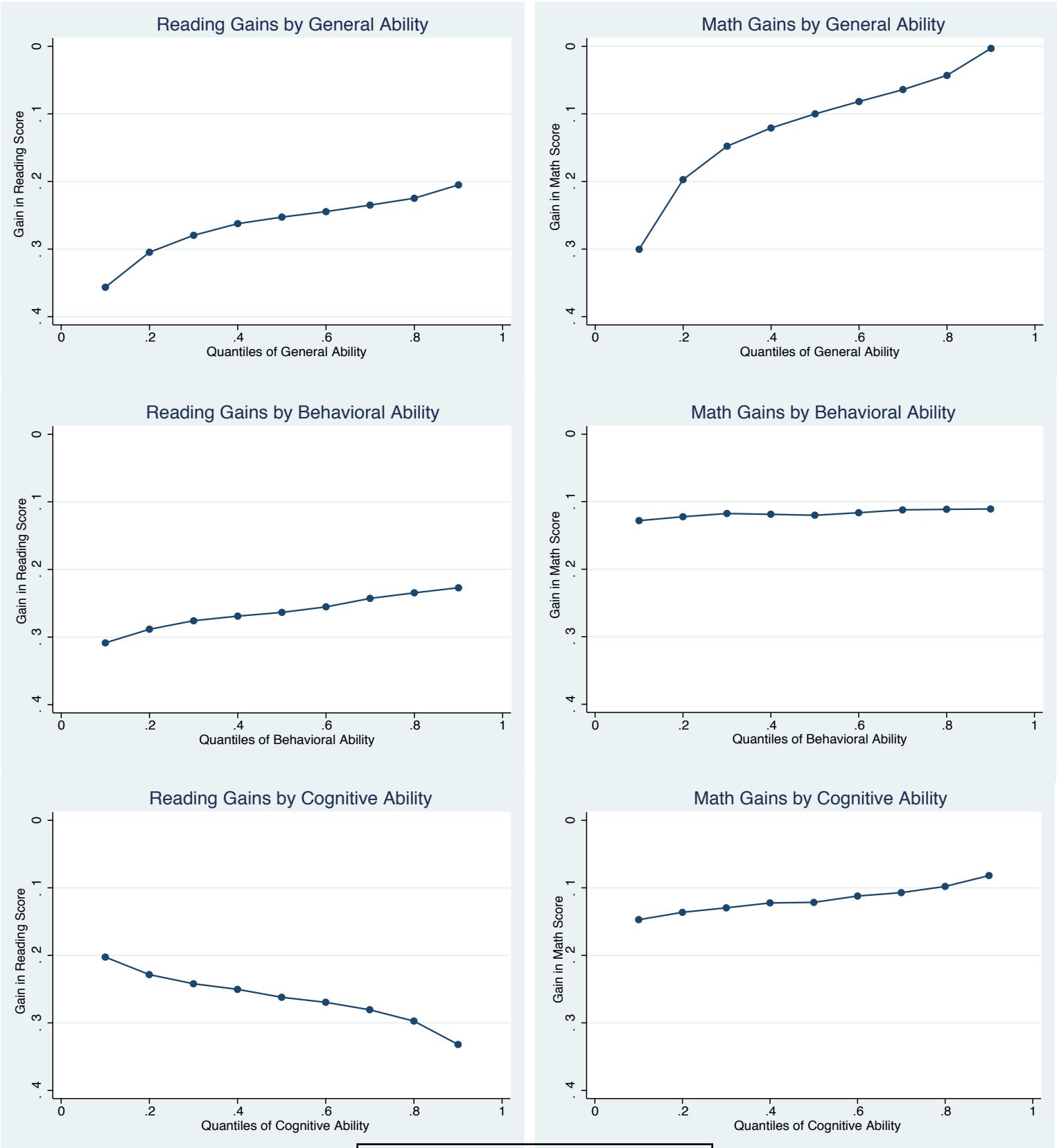
Note: Standard errors are obtained via 1000 bootstrap replications.

Figure D-1: Achievement Gains in 2001/02 by Ability Quantiles



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

Figure D-2: Achievement Gains in 1998/99 by Ability Quantiles



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

E Identification Proofs (Online Appendix, Not For Publication)

Consider a version of the model of equations (5) and (6) in a multiperiod setting. In period 1 the model is given by:

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

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

where these equations are assumed to be free of selection. Without loss of generality, we impose the following normalizations $\alpha_{\zeta,1,1} = 1$, $\pi_{\zeta,1,1} = 1$ and $\phi_{\beta,1,1} = 1$.[48]

Moving forward in time we have that the demeaned selection corrected period t cognitive tests for retention status r are written as

$$\zeta_{i,j,r,t} = A_i \alpha_{\zeta,j,r,t} + B_i \phi_{\zeta,j,r,t} + C_i \pi_{\zeta,j,r,t} + \sum_{\tau=2}^{t} \eta_i^{(\tau)} \delta_{\zeta,j,r,t}^{(\tau)} + \varepsilon_{i,\zeta,j,t}. \tag{10}$$

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

E.1 Asymmetric Factor Distributions

First consider the case in which the distributions of all factors are asymmetric. In this case, we can identify the model with access to less equations than what is assumed in standard factor analysis. In particular suppose that $N_\zeta \geq 2$ so we have access to at least 2 cognitive measures and $N_\beta \geq 2$ so we have access to at least 2 behavioral measures.

We first take cross moments between the j^{th} cognitive and the k^{th} behavioral measure

[48] 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.

for the period 1 selection free outcomes

$$\begin{array}{l} E\left(\left(\zeta_{j,1}\right)^{2}\beta_{k,1}\right)=\alpha_{\zeta,j,1}^{2}\alpha_{\beta,k,1}E\left(A^{3}\right) \\ E\left(\zeta_{j,1}\left(\beta_{k,1}\right)^{2}\right)=\alpha_{\zeta,j,1}\alpha_{\beta,k,1}^{2}E\left(A^{3}\right) \end{array}. \tag{11}$$

Since $E\left(A^{3}\right)\neq 0$, we can form

$$\frac{E\left(\zeta_{j,1}\left(\beta_{k,1}\right)^{2}\right)}{E\left(\zeta_{1,1}\left(\beta_{k,1}\right)^{2}\right)}=\frac{\alpha_{\zeta,j,1}\alpha_{\beta,k,1}^{2}E\left(A^{3}\right)}{\alpha_{\beta,k,1}^{2}E\left(A^{3}\right)}=\alpha_{\zeta,j,1}$$

to recover all of the period 1 general ability loadings on cognitive tests, $\alpha_{\zeta,j,1}$, for $j=2,\ldots,N_{\zeta}$. We can then, for example, form

$$\frac{E\left(\zeta_{1,1}\left(\beta_{k,1}\right)^{2}\right)}{E\left(\left(\zeta_{1,1}\right)^{2}\beta_{k,1}\right)}=\frac{\alpha_{\beta,k,1}^{2}E\left(A^{3}\right)}{\alpha_{\beta,k,1}E\left(A^{3}\right)}=\alpha_{\beta,k,1}$$

and recover the general ability loadings on period 1 behavioral tests.

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

$$\frac{\zeta_{i,j,1}}{\alpha_{\zeta,j,1}}=\left[C_{i}\frac{\pi_{\zeta,j,1}}{\alpha_{\zeta,j,1}}+\frac{\varepsilon_{i,\zeta,j,1}}{\alpha_{\zeta,j,1}}\right]+A_{i},$$

$$\frac{\beta_{i,k,1}}{\alpha_{\beta,k,1}}=\left[B_{i}\frac{\phi_{\beta,k,1}}{\alpha_{\beta,k,1}}+\frac{\varepsilon_{i,\beta,k,1}}{\alpha_{\beta,k,1}}\right]+A_{i}.$$

Then, using Kotlarski (1967), the distribution of A (and of $\left[C\frac{\pi_{\zeta,j,1}}{\alpha_{\zeta,j,1}}+\frac{\varepsilon_{\zeta,j,1}}{\alpha_{\zeta,j,1}}\right]$ and $\left[B\frac{\phi_{\beta,k,1}}{\alpha_{\beta,k,1}}+\frac{\varepsilon_{\beta,k,1}}{\alpha_{\beta,k,1}}\right]$) is nonparametrically identified.

With all of the period 1 parameters associated with general ability A as well as its distribution identified, we can then take the period 1 system of cognitive tests and form

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

for any $j\neq k$ with $j,k=1,...,N_{\zeta}$. By forming

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

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

applying the same argument to the system of behavioral tests, we can recover $\phi_{\beta,j,1}$ and the nonparametric distributions of B and $\varepsilon_{\beta,j,1}$ for $j = 1, ..., N_\beta$.

Once we have recovered the distribution of (A_i, B_i, C_i), we can proceed to the next period. In particular, because we now know the distributions of (A_i, B_i, C_i) ahead of time, we can simply estimate the selection equation by itself and from it recover the pattern of selection for period 2 outcomes. Hence, we can use it to correct the test scores in period 2 for selection.

Now consider identification of equation (10) in period 2 for an arbitrary retention status r. We can form cross second moments between period 2 and period 1 cognitive tests:

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

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

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

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

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

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

to identify the loadings on the permanent shock for all cognitive scores $j = 1, ..., N_\zeta$ and retention status $r = \infty$.[49] We can then apply Kotlarski to any pair of equations j, k for $r = \infty$ and identify the nonparametric distributions of $\eta^{(2)}$ and $\varepsilon_{\zeta,j,2}, \varepsilon_{\zeta,k,2}$. To identify the

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

loadings for retention statuses $r \neq \infty$, we can form

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

Applying the same arguments recursively, it is clear that we can add a new permanent shock every period and still identify all of the loadings and nonparametric distributions of the unobservables.

E.2 Some Symmetric Factor Distributions

Assume that the distribution of general ability, A, is symmetric. Further assume that either the distribution of behavioral ability, B, or the distribution of cognitive ability, C, is not symmetric.[50] In this case, the above strategy (using equation (11)) does not work. Without loss of generality, take the case in which C is assumed to be non-symmetric and assume, as in standard factor analysis, that $N_\beta \geq 3$ and $N_\zeta \geq 3$.

Now, take any two period 1 selection-free cognitive scores and form:

$$\begin{aligned} E\left(\left(\zeta_{j,1}\right)^2\zeta_{k,1}\right) &= \alpha_{\zeta,j,1}^2\alpha_{\zeta,k,1}E\left(A^3\right)+\pi_{\zeta,j,1}^2\pi_{\zeta,k,1}E\left(C^3\right) \\ &= \pi_{\zeta,j,1}^2\pi_{\zeta,k,1}E\left(C^3\right). \end{aligned}$$

where the second line follows from the assumed symmetry of A. Then, by taking ratios:

$$\frac{E\left(\zeta_{1,1}\left(\zeta_{k,1}\right)^2\right)}{E\left(\left(\zeta_{1,1}\right)^2\zeta_{k,1}\right)} = \frac{\pi_{\zeta,k,1}^2 E\left(C^3\right)}{\pi_{\zeta,k,1}E\left(C^3\right)} = \pi_{\zeta,k,1}.$$

we recover the period 1 cognitive ability loadings $\pi_{\zeta,k,1}$ for $k = 2, ..., N_\zeta$.

In order to recover the period 1 general ability loadings on cognitive tests, $\alpha_{\zeta,j,1}$, for $j = 2, ..., N_\zeta$ we can also form

$$E\left(\zeta_{j,1}\beta_{k,1}\right) = \alpha_{\zeta,j,1}\alpha_{\beta,k,1}E\left(A^2\right) \tag{12}$$

and use

$$\frac{E\left(\zeta_{j,1}\beta_{k,1}\right)}{E\left(\zeta_{1,1}\beta_{k,1}\right)} = \frac{\alpha_{\zeta,j,1}\alpha_{\beta,k,1}E\left(A^2\right)}{\alpha_{\beta,k,1}E\left(A^2\right)} = \alpha_{\zeta,j,1}$$

[50] That is, if B (alternatively C) is asymmetric, C (alternatively B) can (but is not required to) be symmetric too.

to identify them. Then, for any pair of period 1 cognitive scores j and k ($j \neq k$), we can calculate cross moments to obtain

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

Assuming $\pi_{\zeta,j,1}\alpha_{\zeta,k,1} - \alpha_{\zeta,j,1}\pi_{\zeta,k,1} \neq 0$, this system of equations gives $E\left(A^2\right)$ and $E\left(C^2\right)$. Then, we can use (12) to recover $\alpha_{\beta,k,1}$ for $k = 1, ..., N_\beta$.

Since we now know all of the period 1 loadings on general ability, we can use the Kotlarski argument to recover the distribution of A (and of $\left[C\frac{\pi_{\zeta,j,1}}{\alpha_{\zeta,j,1}} + \frac{\varepsilon_{\zeta,j,1}}{\alpha_{\zeta,j,1}}\right]$ and $\left[B\frac{\phi_{\beta,k,1}}{\alpha_{\beta,k,1}} + \frac{\varepsilon_{\beta,k,1}}{\alpha_{\beta,k,1}}\right]$) nonparametrically. Next, for any pair of period 1 behavioral scores j and k ($j \neq k$, $j \neq 1$, and $k \neq 1$), we calculate cross moments to obtain the following system of equations:

$$\begin{aligned} E\left(\beta_{1,1}\beta_{j,1}\right) &= \alpha_{\beta,1,1}\alpha_{\beta,j,1}E\left(A^2\right) + \phi_{\beta,j,1}E\left(B^2\right) \\ E\left(\beta_{1,1}\beta_{k,1}\right) &= \alpha_{\beta,1,1}\alpha_{\beta,k,1}E\left(A^2\right) + \phi_{\beta,k,1}E\left(B^2\right) \\ E\left(\beta_{j,1}\beta_{k,1}\right) &= \alpha_{\beta,j,1}\alpha_{\beta,k,1}E\left(A^2\right) + \phi_{\beta,j,1}\phi_{\beta,k,1}E\left(B^2\right). \end{aligned}$$

This gives

$$\begin{aligned} \phi_{\beta,j,1} &= \frac{E\left(\beta_{j,1}\beta_{k,1}\right) - \alpha_{\beta,j,1}\alpha_{\beta,k,1}E\left(A^2\right)}{E\left(\beta_{1,1}\beta_{k,1}\right) - \alpha_{\beta,1,1}\alpha_{\beta,k,1}E\left(A^2\right)} \\ \phi_{\beta,k,1} &= \frac{E\left(\beta_{j,1}\beta_{k,1}\right) - \alpha_{\beta,j,1}\alpha_{\beta,k,1}E\left(A^2\right)}{E\left(\beta_{1,1}\beta_{j,1}\right) - \alpha_{\beta,1,1}\alpha_{\beta,j,1}E\left(A^2\right)} \end{aligned}$$

provided that both denominators are non-zero. Thus, we have obtained $\phi_{\beta,j,1}$ for $j = 2, ..., N_\beta$. By iteratively applying the Kotlarski argument, we can nonparametrically recover the distributions of B and $\varepsilon_{\beta,j,1}$ for $j = 1, ..., N_\beta$. Identification of other parts of the model follows in the same way as before.

Note that we have exploited the assumption $E\left(C^3\right) \neq 0$. If $E\left(C^3\right) = 0$ and $E\left(B^3\right) \neq 0$, then we can still use the same strategy to identify the model. That is, as long as one of the distributions of ability is skewed, our measurements provide enough information to identify all the distributions of factors. For application purposes, we can test if $E\left(A^3\right) = E\left(C^3\right) = 0$ by checking if $E\left(\zeta_{j,1}^2\zeta_{k,1}\right) = 0$ or not. Similarly, we can test if $E\left(A^3\right) = E\left(B^3\right) = 0$.

E.3 All Symmetric Distributions

The problem arises when $E(A^3) = E(B^3) = E(C^3) = 0$. In this case, we need to rely on higher moments and one additional assumption stated below. Consider $\zeta_{1,1}$ and $\beta_{1,1}$. Using

$$\begin{aligned} E\left(\zeta_{1,1}^2\right) &= E\left(A^2\right) + E\left(C^2\right) + E\left(\varepsilon_{\zeta,1,1}^2\right) \\ E\left(\beta_{1,1}^2\right) &= \alpha_{\beta,1,1} E\left(A^2\right) + E\left(B^2\right) + E\left(\varepsilon_{\beta,1,1}^2\right), \end{aligned}$$

fourth order cross moments can be written as

$$\begin{aligned} E\left(\zeta_{1,1}^3 \beta_{1,1}\right) &= \alpha_{\beta,1,1} E\left(A^4\right) + 3\alpha_{\beta,1,1} E\left(A^2\right)\left(E\left(\zeta_{1,1}^2\right) - E\left(A^2\right)\right) \\ E\left(\zeta_{1,1} \beta_{1,1}^3\right) &= \alpha_{\beta,1,1}^3 E\left(A^4\right) + 3\alpha_{\beta,1,1} E\left(A^2\right)\left(E\left(\beta_{1,1}^2\right) - \alpha_{\beta,1,1}^2 E\left(A^2\right)\right). \end{aligned}$$

With additional information $E(\zeta_{1,1}\beta_{1,1}) = \alpha_{\beta,1,1} E(A^2)$, it is straightforward to show that

$$\begin{aligned} E\left(\zeta_{1,1}^3 \beta_{1,1}\right) - 3E\left(\zeta_{1,1}\beta_{1,1}\right) E\left(\zeta_{1,1}^2\right) &= \alpha_{\beta,1,1}\left(E\left(A^4\right) - 3E\left(A^2\right)^2\right) \\ E\left(\zeta_{1,1} \beta_{1,1}^3\right) - 3E\left(\zeta_{1,1}\beta_{1,1}\right) E\left(\beta_{1,1}^2\right) &= \alpha_{\beta,1,1}^3\left(E\left(A^4\right) - 3E\left(A^2\right)^2\right). \end{aligned}$$

Remember that $E(A^4) = 3E(A^2)^2$ for the centered normal distribution. Therefore, if A is not normally distributed (or more generally, the second and fourth moments do not have the same relationship as that of the normal distribution), then we have

$$\alpha_{\beta,1,1}^2 = \frac{E\left(\zeta_{1,1}\beta_{1,1}^3\right) - 3E\left(\zeta_{1,1}\beta_{1,1}\right) E\left(\beta_{1,1}^2\right)}{E\left(\zeta_{1,1}^3\beta_{1,1}\right) - 3E\left(\zeta_{1,1}\beta_{1,1}\right) E\left(\zeta_{1,1}^2\right)}.$$

The sign of $\alpha_{\beta,1,1}$ is given by the sign of $E(\zeta_{1,1}\beta_{1,1})$. Once we recover $\alpha_{\beta,1,1}$, identification of the remaining parts of the model follows in the same way as before (assuming $N_\zeta \geq 3$ and $N_\beta \geq 3$). Intuitively, as long as the second and fourth moments provide sufficiently different information, the model is identified even if all the factors have symmetric distributions.

E.4 References

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