Skill, Standardized Tests, and Fadeout in Educational Interventions*

Elizabeth U. Cascio
Dartmouth College and NBER

Douglas Staiger
Dartmouth College and NBER

May 2, 2011

Abstract

Educational interventions frequently have effects on test scores that fade out over time. The finding is generally interpreted as showing that the cognitive impacts of early intervention are short-lived. We test an alternative hypothesis: that the common practice of rescaling test scores in standard deviation units creates the illusion of fadeout. If a standard deviation in test scores in later grades translates into a larger difference in cognitive skill, an intervention's effect on test scores may fall even as its effect on skill remains constant or rises. We evaluate this hypothesis by fitting the predictions of a model of education production to correlations in standardized test scores across grades and with college-going using both administrative and survey data. Our results imply that up to 20 percent of fadeout is a statistical artifact, arising not from actual decay in cognitive skills but rather from this practice of normalization.

---

* The authors can be contacted by mail at Department of Economics, 6106 Rockefeller Center, Hanover, NH 03755 and by email at elizabeth.u.cascio@dartmouth.edu and at douglas.o.staiger@dartmouth.edu. We thank Kevin Lang and Thomas Lemieux for the conversations that motivated this research. We are also grateful for the outstanding research assistance of Sarah Cohodes and for funding from the National Center for Teacher Effectiveness at the Center for Education Policy Research at Harvard University. Cascio also gratefully acknowledges funding from the National Academy of Education and the National Academy of Education/Spencer Postdoctoral Fellowship Program and a Rockefeller-Haney Fellowship from Dartmouth College. All errors are our own.

I. Introduction

Fadeout in effects on standardized test scores is a pervasive finding from early educational intervention, characterizing programs ranging from compensatory preschools (Almond and Currie, 2010) to class size reduction (Krueger and Whitmore, 2001, Chetty, et al., 2010), as well as teacher assignment (Kane and Staiger, 2008; Jacob, Lefgren, and Sims, 2008; Rothstein, 2010). At the same time, many of these programs have been found to have impacts on longer-term, non-test outcomes, such as educational attainment and earnings. A widely accepted explanation for this pattern of findings is that the impacts of educational intervention on cognitive skills are limited or decay rapidly while their impacts on non-cognitive skills – which are argued to load more heavily onto non-test outcomes – persist (Heckman et al., 2010).[1]

In this paper, we consider an alternative hypothesis: that the common practice of rescaling test scores in standard deviation units creates the illusion of fadeout. The hypothesis is fairly intuitive.[2] If the variance in cognitive skill rises as children progress through school, a standard deviation in test scores in a higher grade, like eighth grade, will map into a greater difference in skill than a standard deviation in test scores in a lower grade, like kindergarten.[3,4] In turn, the magnitude of the relationship between an intervention and standardized test scores may decline even as its impact on skill is unchanging, or even rising.

Evaluating this hypothesis is difficult, however, since skill has no natural metric. Indeed, the near universal practice of normalizing test scores in some way is a practical response to this

[1] We use the term "cognitive skill" rather than "achievement" to be consistent with this discourse.
[2] Surprisingly, this hypothesis has not received much attention in the economics literature, though Lang (2010) recently emphasizes the scaling issue. In the education literature, this issue has received most attention in the context of understanding achievement gaps between groups (see Selzer, Frank, and Bryk (1994), Reardon (2007), and Reardon and Robinson (2007)).
[3] The findings in Chetty et al. (2010) would appear to support this idea: analyzing data from the Tennessee STAR experiment, they find that a percentile point increase in test scores in eighth grade has a larger impact on earnings at age 27 than a percentile point increase in test scores in kindergarten.
[4] The same intuition holds if we were to consider rescaling test scores in percentiles rather than in standard deviation units. We focus on standard deviation units here because the linearity of the underlying transformation considerably simplifies the analysis.

fact, as tests (and thereby test scales) may change as children in a target population age. Given this, our empirical approach works from several patterns that we might expect to see in data if the variance of cognitive skill is increasing across grades. In particular, test scores in consecutive grades should be more correlated with one another at higher grades than at lower grades, since at higher grades, dramatic shifts in the accumulated skill distribution should be less likely. Similarly, test scores in later grades should be more strongly correlated with longer-term outcomes than test scores in earlier grades.

More formally, we lay out a model of education production that draws insight from the empirical literature and parameterizes alternative hypotheses that might generate similar patterns in the data, such as true decay in the effects of educational intervention on cognitive skill and changing test reliability across grades. The variance of accumulated skill in each grade is a function of model parameters. The correlations in test scores across grades and with long-term outcomes are as well, allowing us to estimate the model parameters using minimum distance methods. We use these parameter estimates to predict how the variance in skill evolves across grades, and in turn, the extent to which a rising variance in skill contributes to fadeout. We fit the model using student-level administrative data on standardized test scores and college-going from Charlotte-Mecklenburg Schools and survey data from the National Longitudinal Survey of Youth 1979. Neither data source is ideal for our analysis, but they are complementary and offer independent estimates of the phenomenon of interest.

Our findings suggest that fadeout is in part a statistical artifact. Estimates are remarkably similar across data sets, both suggesting that, for example, by eighth grade, the effects of an intervention in kindergarten on math tests decline by about 12 percent more than the effects on actual math skills. This statistical artifact accounts for roughly 20 percent of total fadeout in the effect of the intervention on test scores, with the remaining 80 percent due to a true decline in

the effects on actual math skills. Notably, we find cognitive skill to be much too persistent to generate the rapid fadeout common to early intervention. Instead, it must be the case that interventions have effects on skill that are more short-lived than seen in our data, or that their effects on cognitive skill are limited or specific to particular domains that are not tested in later grades. We leave it to future research to resolve this issue, but note that it suggests the statistical artifact may matter less for fadeout in practice than our model predicts.

Our findings nevertheless have wider implications for education research. In particular, a byproduct of our analysis is a series of estimates of how the standard deviation in skill changes across grades. These estimates imply, for example, that an intervention in a later grade may have a larger contemporaneous impact on skill than an earlier intervention that has the same impact on test scores, and that gaps in skill grow faster as children progress through school than gaps in standardized test scores would suggest.

II. Education Production

Our objective is to estimate how the variance in skill changes as children progress through school. The challenge for our analysis is that skill itself is unobservable and has no natural metric. To overcome this obstacle, we model the data-generating process for correlations between test scores across grades and with a longer-term, non-test outcome, which can be observed. The parameters of this model imply how the variance of skill evolves across grades and can be estimated using minimum distance methods, as discussed in the section to follow.

A. Model

We begin by modeling the accumulation of cognitive skill, g. Suppose that the skill of child i in grade t is given by:[5]

$$g_{it} = \mu_i + \alpha_{it}, \quad (1)$$

[5] For exposition, we consider one cohort of students which progresses through school one grade per year, so period is synonymous with grade. We discuss our approach to handling grade repetition in the data below.

where μ_i represents *i*'s skill endowment, and α_{it} represents *i*'s cumulative value-added from education as of grade *t*. Cumulative value-added, in turn, is given by:

$$\alpha_{it} = \delta\alpha_{it-1} + v_{it}, \tag{2}$$

where v_{it} is the innovation to skill in grade *t*, and the parameter δ captures the persistence of value-added from previous grades. The model thus allows skill to decay ($0 \le \delta < 1$) or grow ($\delta > 1$) over time. Beyond the non-negativity constraint, however, we place no restrictions on the value of the parameter.

In terms of the statistical properties of the model, we assume that the v_{it} are drawn independently of μ_i and of each other over time. Equation (2) should therefore be thought of as a linear projection, where the innovation is uncorrelated with accumulated achievement by definition. As a result, any factors that may make the value-added of education correlated over time for a given individual, like ability tracking, are subsumed in δ. That is, if students are tracked on the basis of early achievement differences, achievement differences would be more persistent, or the value of δ would be higher.[6]

The variance of these innovations may change across grades, e.g., if changes to skill are larger in early or later grades. We allow the innovation variance to change smoothly across grades according to:

$$\text{var}(v_{it}) = \beta_v \text{ var}(v_{it-1}), \tag{3}$$

where $\beta_v > 0$, and $\beta_v > 1$ $(\beta_v < 1)$ implies higher (lower) innovation variance in later grades.

Test scores, *y*, are assumed to be noisy measures of accumulated skill:

$$y_{it} = g_{it} + \varepsilon_{it}, \tag{4}$$

[6] This is the most appropriate formulation for the present application since early intervention may affect track placement.

where ε_{it} represents a mean zero measurement error drawn independently of skill and independently over time. For our purposes, it is useful to parameterize the error variance, $\sigma^2_{\varepsilon_t}$, using the reliability ratio, λ_{y_t}:

$$\lambda_{y_t} = \frac{\text{var}(g_{it})}{\text{var}(g_{it}) + \sigma^2_{\varepsilon_t}} \tag{5}$$

λ_{y_t} is distinct from what is usually meant by reliability of a test (e.g., test-retest reliability). It is instead the fraction of total variance in test scores accounted for by the variance in accumulated skill. Thus, while λ_{y_t} will capture testing noise, it may also reflect test content or the characteristics of test takers.[7] Because these might change in unpredictable ways, we allow λ_{y_t} to vary across grades and across testing regimes in our estimation.

Finally, we assume that some longer-term non-test outcome – in our case, college going – is a noisy measure of accumulated skill at the end of secondary school, period *T*:

$$w_i = g_{iT} + u_i, \tag{6}$$

where u_i is an independent error term with variance σ^2_u, and without loss of generality w is normalized so that the coefficient on g is equal to one. Again, it is convenient to parameterize the error variance with a reliability ratio:

$$\lambda_w = \frac{\text{var}(g_{iT})}{\text{var}(g_{iT}) + \sigma^2_u} \tag{7}$$

λ_w gives the fraction of the variance in the longer-term outcome explained by final skill. If non-cognitive skills explain more variation in this outcome than in test scores, we would expect to estimate that λ_w is less than λ_{y_t} for all *t*.

[7] Reported test reliabilities would therefore be an upper bound on what we should estimate for this parameter.

Since our findings will ultimately depend on the model we propose, it is important to consider how that model relates to the existing literature. Perhaps the most useful point of reference comes from the literature on teacher value-added. On this front, the value-added specification implied by the model is comparable to one that has recently been validated using experimental data (Kane and Staiger, 2008).[8] To further explore the empirical implications of the model, we compare our model's estimates of key reduced-form parameters – like observed fadeout in the effects of early intervention on test scores (see below) – to the estimates of these parameters elsewhere in the literature.

B. Implications

What are the implications of the model for measurement and interpretation of the impacts of educational intervention? Consider a randomly-assigned intervention in grade *t* that raises test scores in *t* by θ_t standard deviations and in *t+k* by θ_{t+k} standard deviations. That is, $\theta_t \equiv E\left[\tilde{y}_{it,1}\right] - E\left[\tilde{y}_{it,0}\right]$ and $\theta_{t+k} \equiv E\left[\tilde{y}_{it+k,1}\right] - E\left[\tilde{y}_{it+k,0}\right]$, where the subscripts 1 and 0 represent the treatment and control groups, respectively, and $\tilde{y}_{it} = \left(y_{it} - E\left(y_{it}\right)\right) / \sqrt{\text{var}\left(y_{it}\right)}$. As described above, a number of early interventions have found $\theta_{t+k} < \theta_t$, which we term "fadeout." Fadeout is widely interpreted as showing that the effects of educational intervention on cognitive skill decay over time.

[8] In particular, our model implies the following value-added specification for test scores in standard deviation units, $\tilde{y}_{it}$:

$$\tilde{y}_{it} = \left(\frac{\delta a_{t-1} - a_t}{b_t}\right) + \left(\frac{1-\delta}{b_t}\right)\mu_i + \delta\left(\frac{b_{t-1}}{b_t}\right)\tilde{y}_{it-1} + \left(\frac{1}{b_t}\right)v_{it} + \left(\frac{1}{b_t}\right)\left(\varepsilon_{it} - \delta\varepsilon_{it-1}\right)$$

where a_t and b_t represent the mean and standard deviation in period *t* test scores, respectively. Thus, our model implies that period *t* (normalized) test scores are a linear function of fixed student characteristics (captured in μ_i), last period's test score (the coefficient on which captures both true skill decay and changes in the variance of skill), and the current period innovation (which might be modeled with a vector of teacher effects). The coefficients on these variables are time-varying if the standard deviation of test scores changes over time.

By contrast, our model shows that fadeout can stem either from skill depreciation, or from a rising variance in the skill distribution. Under the statistical properties of the model and the assumption of random assignment, θ_t and θ_{t+k} are related by the following expression:

$$\frac{\theta_{t+k}}{\theta_t} = \delta^k \cdot \sqrt{\frac{\text{var}(g_{it})}{\text{var}(g_{i,t+k})}} \tag{8}$$

if test reliability is unchanging across grades.[9,10] The part of fadeout arising from skill depreciation is represented by δ^k, while the part arising from normalization – and so a statistical artifact – is represented by $\sqrt{\text{var}(g_{it})/\text{var}(g_{it+k})}$. The common interpretation of fadeout implicitly assumes that $\sqrt{\text{var}(g_{it})/\text{var}(g_{it+k})} = 1$ for all *t* and *k*. Equation (8) implies, however, that there are scenarios, as when $\delta \geq 1$, where fadeout could be observed without any decay in skill. The equation also shows that normalization of test scores can contribute to fadeout even when $\delta < 1$.

The goal of our empirical analysis is to produce estimates of $\sqrt{\text{var}(g_{it})/\text{var}(g_{it+k})}$ for various values of *t* and *k*, which correspond to the amount of fadeout that is purely a statistical artifact.[11] As noted, the variance of skill cannot be observed, which would generally make this problem intractable. However, the model given above yields an expression for it:

$$\text{var}(g_{it}) = \sigma_\mu^2 + \text{var}(\alpha_{it}), \tag{9a}$$

where σ_μ^2 is again the variance in the skill endowment, and the variance in cumulative value-added is:

---

[9] That is, we assume in the derivation of equation (8) that the tests in *t* and *t*+*k* have the same reliability ratio for all *k*. If the test in *t*+*k* is more (less) reliable than that in *t*, the contribution of test normalization to fadeout will be lower (higher). While we allow the reliability ratio to vary in our estimation, we assume a constant reliability ratio in the calculation of this parameter, since differences in the reliability ratio across grades will be idiosyncratic to any given data set.

[10] Random assignment implies that $E(\varepsilon_{i0})=E(\varepsilon_{i1})$. Below, we discuss scenarios where an intervention's effects on cognitive skill may register as noise if they are transitory. In this case, $E(\varepsilon_{i1})>E(\varepsilon_{i0})$, even with random assignment.

[11] We focus on an intervention in kindergarten, but later interventions could be explored with our estimates.

(9b) $$\text{var}(\alpha_{it}) = \sigma^2_{v_1} \beta_v^{t-1} \sum_{j=0}^{t-1} \delta^{2j} \beta_v^{-j} \ ,$$

where recall that β_v characterizes how the variance in the innovations to skill changes over time, and $\sigma^2_{v_1}$ represents the variance of the first period innovation to skill.

Equations (9a) and (9b) show that the variance of cumulative value-added will be larger and grow faster across grades the larger is δ – or the less skill decays over time – and the larger is β_v – or with persistence or expansion of the innovation variance. The equations also imply that only ratios in the variance of cumulative value-added can be interpreted. For example, we could set the value of either σ^2_{μ} or $\sigma^2_{v_1}$, and the values undertaken by $\sqrt{\text{var}(g_{it}) / \text{var}(g_{it+k})}$ would be unchanged. In practice, we set $\sigma^2_{v_1} = 1$, and σ^2_{μ} is measured in multiples thereof. The next section discusses our approach to estimating these model parameters.

Before turning to this discussion, note that we can put these estimates to other purposes. For example, our model yields a prediction of overall fadeout under the assumption that an intervention has some lasting (though not necessarily completely persistent) impact on cognitive skill (equation (8)), which we can compare to findings from actual educational intervention. Further, we can use estimates of $\sqrt{\text{var}(g_{it}) / \text{var}(g_{it+k})}$ to provide insight into a number of other issues in education research, such as re-interpreting racial and ethnic achievement gaps.

III. Estimation and Identification

Estimation of the policy parameters requires estimation of model parameters. Our estimation strategy takes advantage of the fact that the model characterizes the data generating process for correlations between test scores across grades and between test scores and the non-test outcome. The key insight is that the correlations are observable and unit free. We choose

model parameters to minimize the distance between model predictions and actual correlations observed in the data.

Under the model, this correlation between test scores in periods *t* and *t+k*, regardless of unit of measurement, is given by:[12]

$$\rho_{y_t y_{t+k}} = \sqrt{\lambda_{y_t} \lambda_{y_{t+k}}} \cdot \frac{\sigma_\mu^2 + \delta^k \operatorname{var}(\alpha_{it})}{\sqrt{(\sigma_\mu^2 + \operatorname{var}(\alpha_{it}))(\sigma_\mu^2 + \operatorname{var}(\alpha_{it+k}))}} \tag{10}$$

where $\operatorname{var}(\alpha_{it})$ is given in equation (9b) and all other parameters have been previously defined. Analogously, the correlation between test scores in *t* and the non-test outcome is:

$$\rho_{y_t w} = \sqrt{\lambda_{y_t} \lambda_w} \cdot \frac{\sigma_\mu^2 + \delta^{T-t} \operatorname{var}(\alpha_{it})}{\sqrt{(\sigma_\mu^2 + \operatorname{var}(\alpha_{it}))(\sigma_\mu^2 + \operatorname{var}(\alpha_{iT}))}} \tag{11}$$

Comparison of (11) to (10) shows that the correlation between period *t* test scores and the long-run outcome will be less than the correlation between test scores in period *t* and at the end of secondary school, period *T*, if the long-term outcome is a less reliable measure of period *T* skill.

It is useful to consider several special cases for intuition as to how the model parameters are identified. When $\delta = 0$, or when value-added through education decays immediately, $\rho_{y_t y_{t+k}} = \sqrt{\lambda_{y_t} \lambda_{y_{t+k}}}$ and $\rho_{y_t w} = \sqrt{\lambda_y \lambda_w}$: the correlations depend only on reliability ratios, and skill itself is determined only by child endowments. Under the assumption that all tests are equally reliable, moreover, the correlations between test scores are exactly the same for all *t* and *k*, and the correlations between test scores and long-term outcomes do not vary with *t*.

Continuing with the assumption of constant reliability, consider the case where $\delta = 1$ and $\beta_v = 1$, or where skill accumulation is a random walk and the innovation variance is constant. Because the variance in skill is rising across grades, both $\rho_{y_t y_{t+k}}$ and $\rho_{y_t w}$ fall below the values

[12] We assume in equations (10) and (11) a constant reliability ratio for the purposes of exposition. In our estimation, we allow test reliability to vary across grades and cohorts.

that would be predicted by reliability alone. The correlations also have gradients in *t*, the grade in which the (first) test is administered. In particular, for any given *k*, $\rho_{y_t y_{t+k}}$ is rising in *t*: intuitively, test scores are more correlated in higher grades, when more skill has been accumulated. Similarly, $\rho_{y_t w}$ is greater the later in the school career the test is administered. $\rho_{y_t y_{t+k}}$ is also falling in *k*: the correlations between test scores are weaker the further apart the tests are taken.

To clarify, Figures I and II plot simulated correlations from these special cases. Figure I plots the $\rho_{y_t y_{t+k}}$. The x-axis gives the grade that the first test was administered (*t*). The number of grades ahead the second test is administered (*k*) is represented by the number adjacent to each line; the correlations themselves are represented with solid dots. Thus, the correlation between third and fourth grade test scores is the point on the line labeled with a one (*k*=1) corresponding to grade 3 (*t*=3) on the x-axis. Figure II plots the $\rho_{y_t w}$ and is arranged in a similar fashion, with the grade of the test on the x-axis. We set the reliability ratios and the variance of the skill endowment to be roughly comparable to what we estimate in the CMS data.[13] The figures illustrate the predictions described above: the correlations are unchanging when $\delta = 0$ (Panel A), and increasing in *t* and falling in *k* when $\delta = 1$ and $\beta_v = 1$ (Panel B).

Figures like these also provide a useful vehicle for understanding the implications of further variation in these model parameters. For example, consider a case where $\delta = 1$ and $\beta_v = 0.9$, as shown in Panel C. Here, skill continues to be a random walk, but the innovations to skill are higher variance earlier in a child's school career. In this case, the correlations between test scores strengthen, particularly in higher grades: that is, the $\rho_{y_t y_{t+k}}$ are much more similar

[13] As in our estimation, we also assume that *t*=1 represents kindergarten and *T*=13. We are also setting $\sigma^2_{v_1} = 1$ in these simulations, as we do in our estimation, with σ^2_μ rescaled accordingly.

regardless of *k*, particularly at high values of *t* (Figure I). For example, the correlations between test scores in grades 3 and 4, grades 3 and 5, grades 3 and 6, etc. are all stronger and more similar to each other than was the case in Panel B. Similarly, there is less of a gradient in *t* in the correlation between test scores and the long-term outcome (Figure II). These predictions are intuitive, since more skill is accumulated relatively early. When $\beta_\gamma > 1$, or when skill is accumulated relatively late, the opposite happens: the correlations weaken, and the gradients of the $\rho_{y_t y_{t+k}}$ in *k* and the $\rho_{y_t w}$ in *t* steepen.

Finally, Panel D of each figure shows what happens when $\delta = 0.9$ and $\beta_\gamma = 1$. Here, the variance in value-added remains the same throughout the school career, as was the case in Panel B, but not all of this value-added persists: there is some true fadeout in cognitive skill. The primary consequence of skill decay is to reduce the gradient of $\rho_{y_t y_{t+k}}$ in *t* for any *k*. For example, the correlations between test scores in grades 3 and 4, between grades 4 and 5, between grades 5 and 6, etc. are more similar now than they were in the baseline case. Related, the correlations between test scores and long-term outcome do not rise as quickly in *t*. Again, this is intuitive, since differences in skill accumulated between any two grades will be less than if all skill were to persist. Further, variation in δ appears to add more curvature to these relationships: for example, the correlation between test scores and long-run outcomes – either *w* or much later test scores – can fall across early grades, if skill fades more rapidly than it is accumulating.

Thus, both β_γ and δ affect the speed through which skill is accumulated – albeit through different mechanisms – and these differences manifest in different ways in the data. For example, variation in β_γ is manifested more in the drop off in the test score correlations in *k*, while variation in δ is manifested more in how the $\rho_{y_t y_{t+k}}$ change with *t* for any given *k*. Data on *w* are thus not needed to estimate the model, but long-term outcomes are quite useful for

identification purposes, since β_v and δ are in practice identified only through fairly subtle differences in the patterns of these correlations. In practice, identification of these parameters is complicated further by changing test reliability across grades, which adds noise to the relationships plotted in Figures I and II. We can identify grade-specific reliabilities since we observe multiple correlations for each grade.

IV. Data

Given the discussion above, the ideal data set for estimating the model offers frequent observations over a wide span of grades on a large number of individuals, allowing each correlation to be estimated precisely. Neither of our data sources is ideal in this sense, but the two are complementary and provide independent estimates. This section describes samples and key variables from each and discusses their relative strengths and limitations. In both samples, we focus on math test scores.

A. Data from Charlotte-Mecklenburg Schools

Our administrative data are from Charlotte-Mecklenburg Schools, a large school district in North Carolina. CMS is one of the few school districts in the country with data sufficient for our analysis: similar tests (the North Carolina end of grade (EOG) exams in math) were administered in many consecutive years (1999 to 2009) and consecutive grades (grades 3 to 8), and far enough in the past that we can observe correlations between standardized test scores with a longer-term outcome – whether an individual enrolled in a four-year college in his first year out of high school, collected through the National Student Clearinghouse (NSC) – for several cohorts of students (those in third grade between spring of 1994 and spring of 1998). Another benefit of the CMS data is that the size of the district makes the correlations quite precise, which helps to reduce noise in our estimates. The correlations are estimated on average using about 5400 students per cohort (grade-year).

One drawback of the CMS data, however, is that for no one cohort can we observe all possible correlations. Appendix Table I provides a description of the available data by cohort, where cohorts are defined on the basis of the year (spring) in which an individual should have been in third grade. The table shows that the 1998 cohort is the most recent one for which we are currently able to observe correlations between test scores and college-going. However, the 1998 cohort lacks third grade test score data.[14] Earlier cohorts (1994 to 1997) have tests available in few grades, and later cohorts (1999 to 2008) lack college-going information altogether. In the present analysis, we use all available data, and allow the variance of the endowment (σ_{μ}^{2}) to vary across cohorts in our preferred specifications.

The CMS data present two additional challenges for our analysis. First, the EOG math exams experienced some changes in content and in testing process over the 1999 to 2009 period.[15] In our model, such changes should affect the correlation between test scores administered in separate testing regimes through a change in the test's reliability ratio. Similarly, there were some changes in district demographics over the period, arising both from Hispanic immigration and white return to the district after the end of court-ordered school desegregation in 2001. If test reliability depends on the underlying variance in achievement, these changes could also affect the reliability ratio. To account for all of these factors, we thus allow the reliability ratio (λ_{y}) to vary year-by-year (not just when the tests themselves change), in addition to varying across grades.

Second, our model assumes that year and grade move in lockstep, one-for-one. However, some children repeat grades, while others (albeit a much smaller fraction) skip ahead.

[14] We hope to have college-going information for the 1999 and 2000 cohorts in CMS very soon.
[15] In particular, the math test was changed in 2001, then again in 2006. Changes in the math test editions affected the weights given to different topics. The main change in the testing process over the period was the addition of more time with each edition; the current edition is untimed.

We would like to use these children's scores, but we are uncomfortable assuming, for example, that a child repeating fifth grade and thus administered the fifth grade EOG math exam would occupy the same place in the distribution of sixth grade EOG math score. To confront this complication, we estimated our model using the subset of students who never repeated (or skipped) a grade. This uses a consistent sample to calculate all correlations, but one which is relatively small and less representative of the typical CMS student.

The CMS data present several additional challenges that we are unable to address though modifications to the model or sample. As discussed above, identification is facilitated by having test scores from a wide span of grades. That test score data are available for CMS over such a short grade span thus makes it more difficult to identify the parameters of the model. Second, estimates based on CMS students may not be more widely applicable. To address these issues, we turn to data from the Child and Young Adult component of National Longitudinal Survey of Youth 1979 (NLSY79).

B. Data from the National Longitudinal Survey of Youth

The NLSY79 is a widely-used, nationally representative longitudinal survey of roughly 12,000 individuals who were between the ages of 14 and 21 in 1979. In 1986, the survey incorporated data on the children born to female respondents to the NLSY79, and the children have been surveyed every two years ever since, with most recent data from 2008. The resulting data set – the NLSY79 Child and Young Adult survey – includes biannual scores on Peabody Individual Achievement Tests (PIAT) of math while a child is between the ages of 5 and 14, and information on whether he or she attended college after high school. For the latter, we rely on questions on educational attainment fielded in the young adult questionnaire, which began in 1994, in the survey years when an individual would have turned 18 and 20 or 19 and 21, depending on the cohort.

As noted, one benefit of these data over the CMS data is that test scores are available over a longer span. It is also the case that we can observe all possible correlations between test scores and between test scores and college going for a number of cohorts. We show available data by cohort in Appendix Table II; cohort now corresponds to birth year. We restrict our analysis sample to individuals born between 1982 and 1987, for which we have the complete span of test scores and the information on college-going after high school described above. Although individuals born in 1981 and prior satisfy these same criteria, we exclude them because they are born to increasingly younger mothers given the survey design. As is the case in our analysis of the CMS data, we take account of cohort differences in skill by allowing the endowment variance (σ_{μ}^{2}) to vary across cohorts.

The NLSY79 has several other features that are useful for our analysis. Compared to the EOG math tests in CMS, test content and scoring were consistent over the period relevant for the cohorts in our sample (1986 to 2002), though there were some changes in test administration and the tested population over this span.[16] Again, we account for the effects of these changes by allowing test reliability to change across grades, though we anticipate loosening the restriction of constant reliability will matter less here than in the CMS analysis. Further, children of the same grade are administered the same test regardless of their grade of enrollment. There is therefore no need to restrict attention to non-repeaters.

The main drawbacks of the NLSY79 data are that the samples used to calculate any given correlation are quite small relative to those available for CMS,[17] and data are collected only every two years, not annually. This added noise to our estimates could potentially outweigh any

[16] Most notably, there was a shift to computer aided test administration in 1994. This apparently increased the proportion of children with valid scores by reducing interviewer error in test administration. There was also a move to testing in the English language only in 2002, and the minority oversample of the NLSY79 was also excluded from the 2000 wave (Center for Human Resource Research, 2009).

[17] On average around 350 observations are used to calculate each correlation – 15 times less than is the case in the CMS data.

increases in precision from having a wider grade span represented in the data. Even if this proves to be case, however, the NLSY79 data are more broadly representative.

V. Estimates of the Model Parameters

Tables I presents equal-weighted minimum distance estimates of the model parameters based on the correlation data derived from CMS and the NLSY79, respectively. Panel A of each table pertains to the estimates based on CMS data, while Panel B corresponds to estimates based on the NLSY79.[18]

To establish a benchmark, we begin with a parsimonious specification. In column (1), we force the innovation variance to be unchanging across grades, or $\beta_v = 1$, and assume that the reliability ratio for the test is unchanging across grades and years. The estimates of δ from this specification are above one in both datasets, though not significantly so in CMS. Recall that when $\delta \geq 1$, all fadeout is a statistical artifact. The specifications in remaining columns of the table examine whether this finding continues to hold when we place fewer restrictions on the model estimated.

Before turning to discussion of these next specifications, it is useful to note that estimates of the reliability ratios, λ_y and λ_w, are very precisely estimated and have magnitudes that line up with expectations. In the models estimated on the correlations for CMS, for instance, the value of λ_y=0.883 is consistent with reported reliabilities on the EOG test for math of 0.85-0.90. The estimate of λ_y=0.704 in the NLSY79 data is at the low end of the range of reliability estimates reported for the PIAT math test, but this may reflect noise that arises from other sources unique to the NLSY79, such as inconsistent child effort in a non-school testing environment.

[18] Heteroskedasticity-robust standard errors are in parentheses.

In both samples, the estimates of λ_w are much lower than estimates of λ_y, implying that accumulated cognitive skill contributes considerably less power in explaining the variation in college-going than the variation in test scores.[19] The differential could be made up by non-cognitive skills, implying that non-cognitive skills are relatively more important in determination of college-going, and by generalization, other longer-term outcomes. However, this does not imply that any persistent impacts of early intervention are working more through non-cognitive channels. Indeed, it is completely consistent with early intervention affecting such outcomes (purely) through cognitive channels, particularly if the cognitive effects persist more than previously thought. The low estimates of λ_w in the NLSY79 are consistent with Neal and Johnson's (1996) estimates that the AFQT test explains roughly 15 percent of variation in earnings. Estimates of λ_w maybe be lower in the NLSY79 because self-reports of college going are more noisy than the administrative data on college going available for CMS students.

In column (2), we allow the reliability ratio for the test to vary across grades and years for the reasons outlined above. While we do not do this completely non-parametrically, we do estimate a fairly unrestrictive specification, allowing λ_y to vary across years without restriction but forcing the differences in reliability across grades to be same over time. This specification improves precision for the CMS sample, as reflected in the lower standard errors on key parameter estimates. This is not the case for the NLSY79, reflecting less of a need to make this adjustment for these data. Estimates of δ in this specification are also lower than those in column (1) for both data sets, suggesting that changes in test reliability were previously loading onto this parameter. In particular, in each data set, the estimated reliabilities of earlier tests (not shown) are lower, which when not accounted for would create the illusion of more skill

[19] Estimates of λ_w may be much lower in the NLSY79 because self-reports of college going are more noisy than the administrative data on college-going available for CMS students.

accumulation across grades. Nevertheless, estimates of δ remain statistically indistinguishable from one in both data sets and thus from the hypothesis that fadeout is an illusion entirely.

In column (3) we allow estimates of σ_{μ}^{2} to vary across cohorts. Like allowing for variation in the λ_{y} across grades and years, this improves model fit and slightly reduces the estimates of δ. For this specification, the estimates of δ from the two datasets are nearly identical, and imply a fadeout in true skill of 6-7 percent per year.

The final column of each panel examines whether this conclusion is changed when we loosen the requirement of constancy in the variances of the innovations to value-added. In the specification in column (4), we find no evidence to suggest that the variance of the innovations changes. The estimates of β_{v} are 0.95 and 0.98 for the CMS and NLSY79 samples, and in neither case can we reject the null that the variance is unchanging ($\beta_{v}=1$). Estimates of δ and λ_{w} are little changed from the more restrictive specification in columns (3).

Moving forward, we consider column (4) – the least restrictive model estimated based on all available correlations – to present our preferred estimates. Despite considerable differences in the nature of the tests and the populations tested, this specification yields remarkably similar parameter estimates in CMS and in the NLSY79.

VI. Estimates of the Policy Parameters

A. Example: An Intervention in Kindergarten

Recall that the main objective of our analysis is to understand how much fadeout in the test score impacts of early intervention is a statistical artifact – arising not from actual skill decay, but rather from a rising variance in skill. As discussed above, our answer to this question will be embodied in our estimates of $\sqrt{\text{var}(g_{it})/\text{var}(g_{it+k})}$. For the purposes of exposition and because it makes for a nice point of reference to existing literature, we focus on fadeout from an

intervention in kindergarten (t=0). Recall that we are also assuming a constant reliability of test scores across grades, since variation in test reliability across grades will be idiosyncratic to the application.[20]

The first column in Figure III plots estimates of $\sqrt{\text{var}(g_{i0})/\text{var}(g_{ik})}$ against k for the CMS (top) and NLSY79 (bottom) samples. The underlying estimates of model parameters correspond to the preferred specification, presented in column (4) of Table I. The solid lines connect the estimates themselves, and the dotted lines represent their 95 percent confidence intervals.[21] The top graph (row) in each figure corresponds to the CMS estimates and the bottom graph corresponds to NLSY79.

To set ideas, consider the estimate of $\sqrt{\text{var}(g_{i0})/\text{var}(g_{i8})}$ for CMS in Figure III – 0.884, with a standard error of 0.041 – in reference to equation (8). This estimate answers the counterfactual question: how much fadeout would be observed by eighth grade if the math skills gained from an intervention in kindergarten persisted in full; that is, what would θ_8/θ_0 be if $\delta = 1$? The answer to this question is: almost 12 percent. While concise, this explanation is a bit misleading, since estimates of $\sqrt{\text{var}(g_{i0})/\text{var}(g_{i8})}$ depend on estimates of δ. So more precisely, the estimate conveys how much more fadeout is observed that would be predicted by true skill decay alone. Put differently yet, observed fadeout (θ_8/θ_0) would have to be scaled up by a factor of 1.13 (=1/0.884) to determine how much math *skill* gained from an intervention in kindergarten persisted through eighth grade.

---

[20] If in general, though, tests become more reliable as children age, the contribution of the statistical artifact to observed fadeout diminishes.

[21] Standard errors were calculated using the delta method. Given that the underlying specification estimated different values of σ_μ^2 for each cohort, we had to choose a cohort at which to evaluate each variance (see equation (9a)). We chose the 1998 cohort for the CMS estimates and the 1986 cohort for the NLSY79. The substantive conclusions are unchanged if we were to use different cohorts.

The graphs demonstrate the variation in estimates of this parameter as we change k and the underlying data. In general, the further out in time the second test is given – or the larger is k – the greater the need for rescaling becomes; the variance in skill continues to rise as children progress through school. But the contribution of this phenomenon is greater in earlier grades, when the variance in skill is expanding most quickly. Estimates of the policy parameters are also quite similar across the two data sets, as might have been expected given similarity in the underlying model parameter estimates.

The graphs in the second column of Figure III plot estimates of θ_k/θ_0 (equation (8)) against k, along with the 95 percent confidence intervals. These figures plot the total fadeout of a kindergarten intervention from both skill decay *and* the statistical artifact due to rescaling under the assumption that the intervention had an effect on g. The point estimates are large, exceeding 50 percent after 8 years, but these estimates have wide confidence intervals. The graphs in the final column of Figure III plot the percentage of total fadeout that we estimate is due to statistical artifact, i.e. the decline due to the growing variance of skill seen in the first column of Figure III as a percentage of the total decline seen in the second column. The estimates imply that roughly 20 percent of fadeout is due to the statistical artifact. While the confidence intervals are wide, we can reject the hypothesis that a large fraction of observed fadeout is due to statistical artifact.

B. Comparison to the Literature on Early Intervention

Our estimates of total fadeout can in principle be compared against true estimates of fadeout in the literature on early intervention. One intervention that has tested children annually is the Tennessee STAR experiment, also referred to as Project STAR (Student Teacher Achievement Ratio). Project STAR randomly assigned children entering kindergarten in 1985 to small and regular sizes in roughly 80 Tennessee schools. Children entering these schools in later grades were also randomly assigned to class types, and participants were expected to stay in their

initially assigned type of class through third grade, after which the experiment ended. Krueger and Whitmore (2001) estimate the effects on attending a small class in Project STAR on test scores through eighth grade.

How well do our predictions of total fadeout match up with Krueger and Whitmore's estimates of fadeout? Several issues complicate our attempt to answer this question: we normalize tests in standard deviation units and look at math, while they use percentile ranks and average across performance on math and reading. Further, the treatment was not confined just to kindergarten, but spanned up to four years for some students. These issues aside, their estimates imply that θ_k/θ_0 takes on a value of about 0.2 to 0.3 by eighth grade, which is close to the lower bound of what we estimate for math scores in either of the two data sets.

Nevertheless, our estimates of θ_k/θ_0 for the earliest grades reject the rapid fadeout seen in Project STAR: most of 70 to 80 percent fadeout in effects on test scores in Project STAR occurs immediately after the class size intervention ends. Similarly, estimates of fade-out in teacher effects reach 50 percent or higher only one year out (Kane and Staiger, 2008; Jacob, Lefgren, and Sims, 2008; Rothstein, 2010). Our model does allow for true skill decay of this magnitude through a value of δ well below one. Such a value of δ does not, however, fit the data: cognitive skill appears to be much more persistent.

How can we explain the divergence between our model's predictions of fadeout and those found in the literature? As noted, we assumed in our derivation of overall fadeout in equation (8) that the intervention in question had a true impact on cognitive skill, *g*. For the sake of argument, suppose that the noise in test scores for child *i* at time *t* can be decomposed into some "skill" that fades immediately, τ_{it}, and all remaining sources of error in the test, v_{it}:

$$\varepsilon_{it} = \tau_{it} + v_{it}. \tag{12}$$

In the extreme case where an intervention impacts these other skills τ but not *g*, its effects on test scores will disappear immediately. More realistically, it may be the case that much of the effects of early intervention on test scores arise through effects on τ rather than *g*.

So what might τ in equation (12) represent? There are several candidate interpretations, with very different implications. For example, τ may represent teaching to the test. In this case, rapid fadeout would imply limited or no impacts of an intervention on cognitive skill. But τ may also capture true learning. For instance, while test scores capture general intelligence – and so may be highly correlated from one grade to the next for a given child – standard test instruments are designed to assess current knowledge and understanding. If an intervention primarily affects understanding of material that is not tested in subsequent years, it would manifest as a transitory increase in test performance, even though this understanding might persist in full. Alternatively, interventions may have true effects on cognitive skill that fade rapidly. For example, mixing of treatment and control students in the same classrooms after a classroom-level intervention has ended may promote rapid reversion toward the mean in student performance on material that students learn in school, while student-level test performance may fade less slowly as it also includes more persistent student-level factors.

Gaining a better understanding of why the effects of early intervention fade out so much more rapidly than we would expect is an important topic for future research. Regardless of the explanation, the very existence of more rapid fadeout in practice than applied in our estimates of policy parameters above suggests a more limited role for the statistical artifact: the estimates presented above may be an upper bound.

VII. Conclusion

Educational intervention has frequently been found to have effects on test scores that fade out over time. The finding is generally interpreted as showing that the cognitive impacts of

early intervention are short-lived. We test an alternative hypothesis: that the common practice of measuring test performance in standard deviation units creates the illusion of fadeout. If a standard deviation in test scores in later grades translates into a larger difference in cognitive skill, an intervention's effect on standardized test scores may fall even as its effect on skill remains constant or rises.

We evaluate this hypothesis by fitting the predictions of a model of education production to correlations in standardized test scores across grades and with college-going using both administrative and survey data. Our results imply that up to roughly 20 percent of fadeout is a statistical artifact, arising not from actual decay in cognitive skills but rather from this practice of normalization. While low power makes some of our estimates less informative than we would like, we are generally able to reject the hypothesis that no fadeout is a statistical artifact, however, which has been an assumption implicit in all research on early intervention to date.

Our findings also have wider implications. This is perhaps most notably the case for research on racial, ethnic, and socioeconomic achievement gaps, which are commonly measured in standard deviation units (for a recent overview of this research, see Reardon and Robinson (2007)). Our findings suggest a standard deviation difference in test scores translates into a somewhat larger difference in skill in higher grades, with the implication that gaps in skill expand more rapidly as children progress through school than any expansion in the test score gap would imply. Further, if the variance of skill grows in age, two interventions that have identical impacts on standardized test scores but are administered in different grades are not equivalent. This has potentially important implications for comparing the benefits of interventions that take place at different points in the school career. Fuller exploration of these issues is an important topic for future research.

References

Almond, Douglas and Janet Currie. 2010. "Human Capital Development Before Age Five." *Handbook of Labor Economics*, Volume 4, forthcoming.

Center for Human Resource Research. 2009. *NLSY79 Child & Young Adult Data User Guide: A Guide to the 1986-2006 Child Data 1994-2006 Young Adult Data.* Columbus, Ohio: The Ohio State University.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2010. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." NBER Working Paper 16381.

Heckman, James J., Lena Malofeeva, Rodrigo Pinto, and Peter A. Savelyev. 2010. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." Unpublished Manuscript, University of Chicago.

Jacob, Brian, Lars Lefgren, and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources*, forthcoming.

Kane, Thomas J. and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Mimeo, Harvard University and Dartmouth College.

Krueger, Alan and Diane Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111: 1-28.

Lang, Kevin. 2010. "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member." *Journal of Economic Perspectives* 24(3): 167-182

Reardon, Sean F. 2007. "Thirteen Ways of Looking at the Black-White Test Score Gap." Mimeo, Stanford University.

Reardon, Sean F., & Robinson, J.P. 2007. "Patterns and trends in racial/ethnic and socioeconomic achievement gaps." In Helen A. Ladd & Edward B. Fiske (Eds.), *Handbook of Research in Education Finance and Policy*. Lawrence Erlbaum.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1): 175-214.

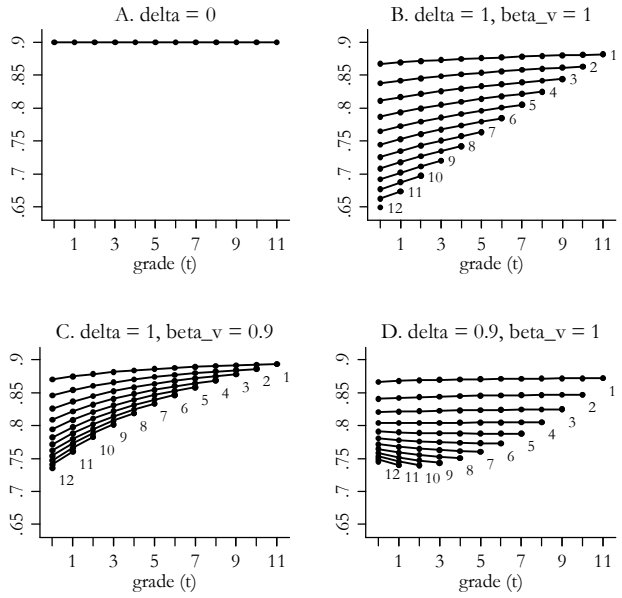
Selzer, Michael H., Ken A. Frank, and Anthony S. Bryk. 1994. "The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric." *Educational Evaluation and Policy Analysis* 16:41-49.

Table I. Estimates of Model Parameters

	A. Charlotte-Mecklenburg Schools Data				B. NLSY79 Child and Young Adult Data			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
σ_μ^2 (natural log)	2.424	2.447			-20.120	2.048		
	(0.152)	(0.065)			(0.000)	(0.360)		
δ	1.06	0.996	0.936	0.922	1.164	0.960	0.935	0.933
	(0.035)	(0.031)	(0.033)	(0.040)	(0.024)	(0.062)	(0.052)	(0.057)
λ_w	0.289	0.29	0.283	0.276	0.167	0.203	0.206	0.203
	(0.006)	(0.005)	(0.006)	(0.009)	(0.012)	(0.017)	(0.016)	(0.023)
λ_y	0.883				0.704			
	(0.005)				(0.014)			
β_v				0.947				0.985
				(0.059)				(0.097)
Root MSE	0.0150	0.0104	0.00793	0.00795	0.0159	0.00888	0.00752	0.00751
Observations	145	145	145	145	90	90	90	90
Model:								
λ varies grade x year		X	X	X		X	X	X
σ_μ^2 varies by cohort			X	X			X	X

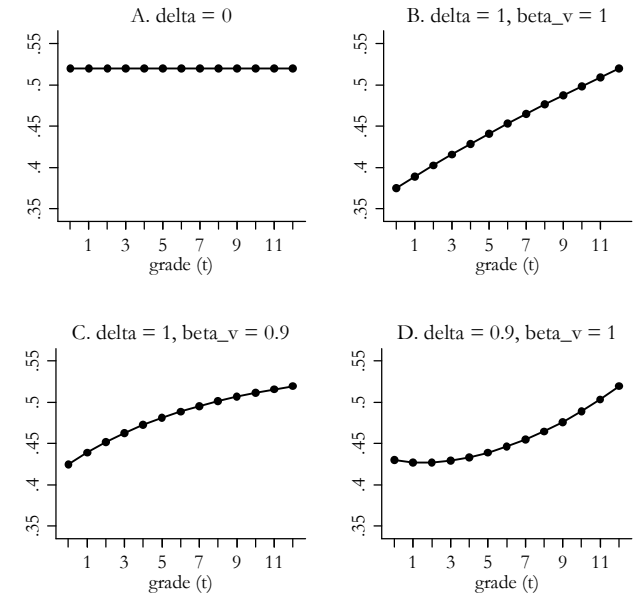
Note: The models are estimated using equal weighted minimum distance. See text for description of the model and the data. Robust standard errors are in parentheses.

Figure I. Simulations of corr(y(t),y(t+k))
Under Alternative Assumptions on Model Parameters



Note: The numbers plotted represent k. Setting lambda=.9, lambda_w=0.3, and sig2_mu=12.

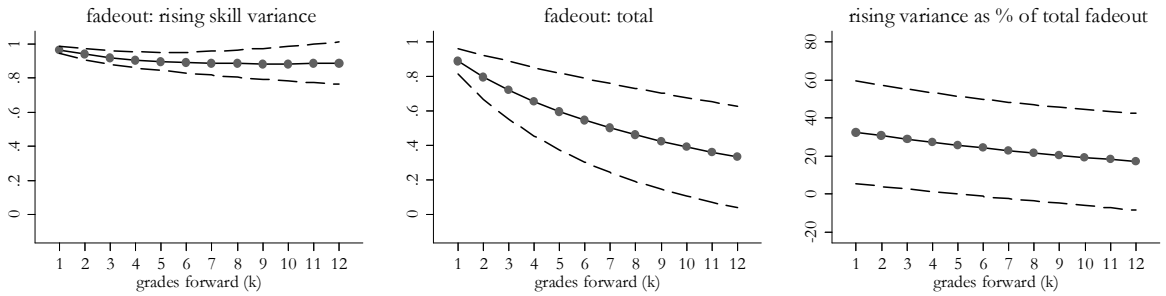
Figure II. Simulations of corr(y(t),w)
Under Alternative Assumptions on Model Parameters



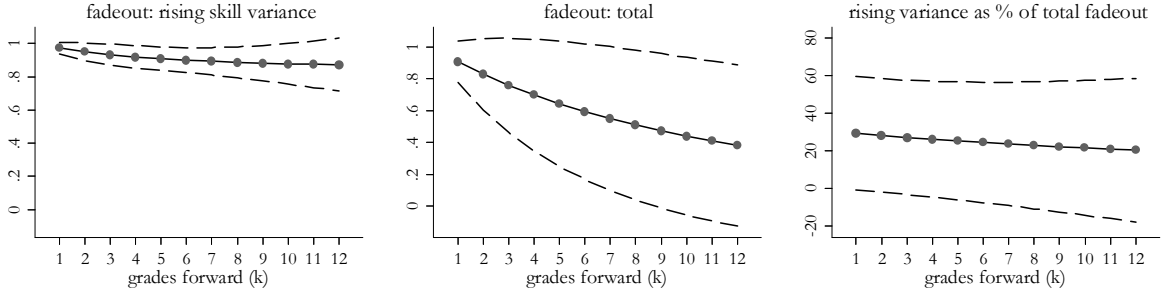
Note: Setting lambda=.9, lambda_w=0.3, and sig2_mu=12.

Figure III. Estimates of Policy Parameters: Intervention in Kindergarten

Charlotte-Mecklenburg Schools



National Longitudinal Survey of Youth - Child/Mother



Note: Panels labeled "fadeout: rising skill variance" plot estimates of $\sqrt{\text{var}(g_{i0})/\text{var}(g_{ik})}$ against *k*. Panels labeled "fadeout: total" plot estimates of θ_k/θ_0 against *k*. Panels labeled "rising variance as percent of total" plot estimates of the fraction of total fadeout accounted for by rising skill variance, $(1-\sqrt{\text{var}(g_{i0})/\text{var}(g_{ik})})/(1-\theta_k/\theta_0)$. The underlying specification is that presented in column (4) of Table I. The estimates shown use variance estimates for the 1998 cohort in CMS (row 1) and the 1986 birth cohort in the NLSY-CM (row 2).

Appendix Table I. Data Availability by Cohort: Administrative Data from Charlotte-Mecklenburg Schools

Year (Spring) in Grade:						
3	4	5	6	7	8	College 1
1994	1995	1996	1997	1998	1999	2004
1995	1996	1997	1998	1999	2000	2005
1996	1997	1998	1999	2000	2001	2006
1997	1998	1999	2000	2001	2002	2007
1998	1999	2000	2001	2002	2003	2008
1999	2000	2001	2002	2003	2004	2009
2000	2001	2002	2003	2004	2005	2010
2001	2002	2003	2004	2005	2006	2011
2002	2003	2004	2005	2006	2007	2012
2003	2004	2005	2006	2007	2008	2013
2004	2005	2006	2007	2008	2009	2014
2005	2006	2007	2008	2009	2010	2015
2006	2007	2008	2009	2010	2011	2016
2007	2008	2009	2010	2011	2012	2017
2008	2009	2010	2011	2012	2013	2018

Notes: The shaded cells correspond to years in which data are available for CMS. The test score data thus span 1999 2009, and college-going in the first year after high school is available through fall 2008. The present analysis data on all cohorts.

Appendix Table II. Data Availability by Cohort: National Longitudinal Survey of Youth 1979 Child-Young Adult Survey

	Year Turn Age:										College Going:	
Year Born	5	6	7	8	9	10	11	12	13	14	Year 1	Year 2
1972	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1991	1993
1973	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1992	1994
1974	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1993	1995
1975	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1994	1996
1976	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1995	1997
1977	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1996	1998
1978	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1997	1999
1979	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1998	2000
1980	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1999	2001
1981	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	2000	2002
1982	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	2001	2003
1983	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	2002	2004
1984	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	2003	2005
1985	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2004	2006
1986	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2005	2007
1987	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2006	2008
1988	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2007	2009
1989	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2008	2010
1990	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2009	2011
1991	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2010	2012
1992	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2011	2013
1993	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2012	2014
1994	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2013	2015
1995	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2014	2016
1996	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2015	2017
1997	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2016	2018
1998	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2017	2019
1999	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2018	2020
2000	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2019	2021
2001	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2020	2022

Notes: The shaded cells correspond to years in which either PIAT assessments were administered in the child survey or college-going information is reported in young adult survey, The child survey began in 1986 and the young adult survey began in 1994; both are administered biennially. The present analysis uses data on the 1982 to 1987 birth cohorts.