NEW EVIDENCE ON THE LONG-TERM IMPACTS OF TAX CREDITS[1]

Raj Chetty, Harvard University and NBER
John N. Friedman, Harvard University and NBER
Jonah Rockoff, Columbia University and NBER

November 2011

ABSTRACT

An important rationale for tax credits is to improve opportunities for children born to low-income families. We combine data on children in a large urban school district with administrative tax records to estimate the impact of tax credits on children's future earnings and other long-term outcomes. Our analysis consists of two parts. We first identify the impacts of tax credits on test scores using non-linearities in tax credits as well as time-series variation in program generosity. We find that a $1,000 increase in tax credits raises students' test scores by 6% of a standard deviation, using our most conservative specification. We then examine the implications of these score gains for earnings using assignment to teachers as an instrument for score. We show that higher scores increase students' probability of college attendance, raise earnings, reduce teenage birth rates, and improve the quality of the neighborhood in which their students live in adulthood. Our results suggest that a substantial fraction of the cost of tax credits may be offset by earnings gains in the long run.

I. Introduction

Many tax policy provisions direct subsidies to low income families with children under the belief that such redistribution not only helps these households' immediate circumstances but also provides for better opportunities for children. Yet despite the centrality of this belief to tax policy, to our knowledge no paper has examined this very important topic. This hole in the literature reflects the difficulty of studying the long-term impacts of these tax programs. In this paper, we provide the first evidence on this issue by focusing on one particular long-run channel: education. While taxes could have long-term impacts in many ways, education is thought to be a particularly important pathway for such effects. One practical advantage of studying the education channel is that test scores provide excellent short-term metrics of progress for nearly all children. Furthermore, previous research suggests that test scores provide a good proxy for the long-term outcomes of children as young as 5 (e.g., Chetty et al., forthcoming; Heckman et al., 2010c; Kane and Staiger, 2008). If cash transfers increase test scores, and those increases have a causal impact on adult outcomes, then examining the impact of tax credits on children's test scores provides an easy way to examine the long-run impact of income on children without needing to observe children for many years into adulthood.

Two recent papers have examined the short-term impacts of tax credits on test scores, but these papers have reached conflicting conclusions. Dahl and Lochner (2011) find large but imprecisely estimated effects on children's test scores from the increase in the Earned Income Tax Credit between 1994 and 1996. In contrast, Jacob and Ludwig (2007) find a precisely estimated zero effect from the outcome of a randomized housing subsidy lottery in Chicago. Neither of these studies has attempted to quantify the impact of test score improvements on long-term outcomes such as earnings, which is our focus here.

We analyze the impacts of taxes on educational achievement and earnings by combining two datasets to form a large sample linking student educational records with family background and income. The first dataset contains administrative data from a large urban school district. These data include information on the test scores of children in grades 3–8. The data also include a rich set of individual characteristics, including age, gender, race, ethnic background, and English proficiency. The second dataset includes selected data from U.S. tax records for all families in the school district sample, as well as adult outcomes for the students themselves (when old enough). These data provide important family background characteristics, such as income and marital status, often missing from other administrative educational datasets. In addition, these data provide precise information on the eligibility of families for various federal and state credits such as the Earned Income Tax Credit (EITC) and Child Tax Credit (CTC).

The ideal empirical setup for this question would link quasi-experimental variation in the receipt of tax credits across families directly with long-term data following those children into adulthood. Unfortunately, the electronic U.S. tax records do not cover enough years to perform this analysis. Instead, we conduct the analysis of the long-term impacts of tax credits in two parts. In the first part of the paper, we identify the impact of cash transfers on children's test scores. We first exploit the non-linearities in the transfer schedule present in the EITC and CTC. By controlling for a flexible function between family income and educational achievement that is smooth across all income ranges, we isolate the sharply non-linear variation in income from tax credits. The second strategy uses changes in transfer policies over time. The primary policy we examine here is the Earned Income Tax Credit. While the federal credit has remained stable since 1996, there have been a series of increases in state and local match rates for the federal EITC. We exploit these increases in the credit for a second source of identification that is orthogonal to the first approach.

Using both empirical approaches, we find a large and precisely estimated relationship between cash transfers and student test scores. Our estimates imply that each $1,000 in tax credit increases student test scores by 6–9% of a standard deviation (SD). These effects are larger in math (9.3%) than in reading (6.2%) and are larger for students in middle school (8.5%) than in elementary school (7.3%). The size of these data results in standard errors on the order of 0.5% of a standard deviation, and so all of these differences are highly statistically significant. The results are robust to a number of different flexible sets of controls. Consistent with the time-series policy variation, the impact of federal credits per se increase sharply over time.

In the second part of the paper, we directly estimate the causal link between scores and adult outcomes of students using student-teacher links as an instrument for test scores. We begin by presenting evidence that teacher assignment is a valid instrument for scores. In particular, we show that selection on observables is minimal by establishing that parent characteristics are uncorrelated with teacher assignment. To evaluate sorting on unobservables, we use a quasi-experimental method of testing for bias in teacher assignment that exploits changes in teaching assignments at the school-grade level, following Chetty, Friedman, and Rockoff (2011). We find that the predicted impacts closely match observed impacts, suggesting that bias due to selection on unobservables is minimal.

Having established that teacher assignment is a plausible instrument for test scores, we then analyze the causal impact of score increases. We find that scores have substantial impacts on a broad range of outcomes. A 1 SD improvement in scores in a single grade raises the probability of college attendance at age 20 by about 5 percentage points, relative to a sample mean of 37%. Improvements in scores also raise the quality of the colleges that students attend, as measured by the average earnings of previous graduates of that college. Students who score higher have steeper earnings trajectories, with significantly higher earnings growth rates in their 20s. At age 28, the oldest age at which we have a sufficiently large sample size to estimate earnings impacts, a 1 SD increase in scores in a single grade raises earnings by almost 9% on average. Assuming that this 9% impact on earnings remains constant over time, the mean student would gain more than $40,000 in lifetime income (with a 3% discount rate) from a 1 SD improvement in scores in a single year. We also find that improvements in scores significantly reduce the probability of having a teenage birth, increase the quality of the neighborhood in which the student lives (as measured by the percentage of college graduates in that ZIP code) in adulthood, and raise 401(k) retirement savings rates. The impacts on adult outcomes are all highly statistically significant, with the null of no impact rejected with $p < 0.01$.

Combining our estimates of the impacts of tax credits on scores and scores on earnings, we find that each dollar of income through tax credits increases net present value (NPV) earnings by more than $1. These results suggest that a substantial fraction of the cost of tax credits may be offset by earnings gains in the long run. Hence, when analyzing the costs and benefits of policies such as the Earned Income or Child Tax Credit, policy makers should carefully consider the potential impacts of these programs on future generations.

The remainder of this paper is organized as follows. In Section II, we describe the data used in our analysis. Section III describes the specifics of the tax policies we examine, as well as our main findings regarding the impact of income transfers on student test scores. Section IV analyzes the link between scores and adult outcomes using teacher quality as an instrument. Section V combines estimates from the two parts of the paper to calculate the impact of cash transfers on students' long-run outcomes.

II. Data

We draw information from two administrative databases: students' school district records and information on these students and their parents from U.S. tax records. The analysis dataset combines selected variables from individual tax returns, third party reports, and information from the school district database, with individual identifiers removed to protect confidentiality. We first describe the two data sources and then the structure of the linked analysis dataset. Finally, we provide descriptive statistics and cross-sectional correlations using the analysis dataset.

II.A. School District Data

We obtain information on students, including enrollment history, test scores, demographics, and teacher assignments from the administrative records of a large urban school district. These data span the school years 1988–1989 through 2008–2009 and cover roughly 2.5 million children in grades 3–8. For simplicity, we refer below to school years by the year in which the spring term occurs, e.g., the school year 1988–89 is 1989.

Test Scores—The data include approximately 18 million test scores. Test scores are available for English language arts and math for students in grades 3–8 in every year from the spring of 1989 to 2009, with the exception of 7th grade English scores in 2002. In the early and mid 1990s, all tests were specific to the district. Starting at the end of the 1990s, the tests in grades 4 and 8 were administered as part of a statewide testing system, and all tests in grades 3–8 became statewide in 2006 as required under the No Child Left Behind law.[2] Because of this variation in testing regimes, we follow prior work on measuring teachers' effects on student achievement, taking the official scale scores from each exam and normalizing the mean to 0 and the standard deviation to 1 by year and grade. The within-grade variation in achievement in the district we examine is comparable to the within-grade variation nationwide, so that our results can easily be compared to estimates from other samples.[3]

Demographics—The dataset also contains information on ethnicity, gender, age, receipt of special education services, and limited English proficiency for the school years 1989 through 2009. The database used to code special education services and limited English proficiency changed in 1999, creating a break in these series that we account for in our analysis by interacting these two measures with a post-1999 indicator. Information on free and reduced price lunch, received by more than three quarters of the students in the school district, is available starting in school year 1999.

Teachers—The dataset links students to classrooms and teachers for students in grades 3–8 from 1991 through 2009.[4] This information is derived from a data management system which was phased in over the early 1990s, so not all schools are included in the first few years of our sample. In addition, data on course teachers for middle and junior high school students—who, unlike students in elementary schools, are assigned different teachers for math and English—are more limited. Course teacher data are unavailable prior to the school year 1994, then grow in coverage to roughly 60% by school year 1998 and 85% by 2003. Even in the most recent years of the data, some middle and junior high schools do not report course teacher data, and in these years roughly 15% of the district's students in grades 6 to 8 are not linked to math and English teachers.

The missing teacher links raise two potential concerns. First, our estimates (especially for grades 6–8) apply to a subset of schools with more complete information reporting systems and thus may not be representative of the district as a whole. These schools do not differ significantly from the sample as a whole on test scores and other observables. Second, and more importantly, missing data could generate biased estimates. Almost all variation in missing data occurs at the school level because data availability is determined by whether the school utilizes the district's centralized data management system for tracking course enrollment and teacher assignment. Specifications that exploit purely within-school comparisons are therefore largely unaffected by missing data and we show that our results are robust to exploiting such variation. Moreover, we obtain similar results for the subset of years for which we have complete data coverage in grades 3–5, confirming that missing data do not drive our results.

We obtain information on teacher experience from human resource records. The human resource records track teachers from when they started working in the district and hence give us an uncensored measure of within-district experience for the teachers in our sample. However, we lack information on teaching experience outside of the school district.

Sample Restrictions—Starting from the raw dataset, we make the following sample restrictions to obtain our analysis sample. First, because we always condition on prior test scores, we restrict our sample to grades 4–8, where prior test scores are always available. Second, we drop the 2% of observations where the student is listed as receiving instruction at home, in a hospital, or in a school serving solely disabled students. We also drop the 6% of observations for students in classrooms where more than 25% of students are receiving special education services, as these classrooms may be taught by multiple teachers or have other special teaching arrangements. Finally, when a teacher is linked to students in multiple schools during the same year—this occurs in 0.3% of cases—we use only the links for the school where the teacher is listed as working according to human resources records and set the teacher as missing in the other schools. After these restrictions, we are left with 15 million student-year-subject observations, 9.1 million of which have teacher information.

II.B. Tax Data

We obtain data on students' adult outcomes and their parents' characteristics from income tax returns (e.g., form 1040) and third-party reports on wage earnings (form W–2) and college attendance (form 1098–T). We only link students born prior to 1991 to the tax data because the earliest adult outcome that we measure is at age 20.

A detailed description of the dataset and variables is given in Chetty et al. (2011), who use these data to study the long-term impacts of Project STAR. Here, we briefly summarize some key features of the variables used below. The year always refers to the tax year (i.e., the calendar year in which the income is earned or the college expense incurred). In most cases, tax returns for tax year *t* are filed during the calendar year *t* + 1. We express all monetary variables in 2010 dollars, adjusting for inflation using the Consumer Price Index.

Earnings—Individual earnings data come from W–2 forms, which are available from 1999–2010, and cover both tax filers and non-filers. Individuals with no W–2 are coded as having 0 earnings.[5] We cap earnings in each year at $100,000 to reduce the influence of outliers; 1.2% of individuals in the sample report earnings above $100,000 in a given year.

College Attendance—We define college attendance as an indicator for having one or more 1098–T forms filed on one's behalf. Title IV institutions—all colleges and universities as well as vocational schools and other postsecondary institutions—are required to file 1098–T forms that report tuition payments or scholarships received for every student.[6] The 1098–T data are available from 1999–2009. Comparisons to other data sources indicate that 1098–T forms accurately capture U.S. college enrollment.[7] Because the data are based on tuition payments, we have no information about college completion or degree attainment.

College Quality—We construct an earnings-based index of college quality as in Chetty et al. (2011). Using the full population of all individuals in the United States aged 20 on 12/31/1999 and all 1098–T forms for year 1999, we group individuals by the higher education institution they attended in 1999. We take a 0.25% random sample of those not attending a higher education institution in 1999 and pool them together in a separate "no college" category. For each college or university (including the "no college" group), we then compute average W–2 earnings of the students in 2009 when they are aged 30. Among colleges attended by students in our data, the average value of our earnings index is $42,932 for four-year colleges and $28,093 for two-year colleges.[8] For students who did not attend college, the imputed mean wage is $16,361.

Neighborhood Quality—We use data from 1040 forms to identify each household's ZIP code of residence in each year. For non-filers, we use the ZIP code of the address to which the W–2 form was mailed. If an individual was not required to file and has no W–2 in a given year, we impute current ZIP code as the last observed ZIP code. We construct a measure of neighborhood quality using data on the percentage of college graduates in the individual's ZIP code from the 2000 Census.

Teenage Birth—We first identify all women who claim a dependent when filing their taxes at any point before the end of the sample in tax year 2010. We observe dates of birth and death for all dependents and tax filers until the end of 2010 as recorded by the Social Security Administration. We use this information to define a teenage birth as ever claiming a dependent who was born in a year when the mother was age 13–19 as of 12/31. Note that this definition of teenage birth suffers from three sources of measurement error. First, it does not capture teenage births to individuals who *never* file a tax return before 2010. This is relatively rare. Second, the mother must herself claim the child as a dependent at some point during the sample years. If the child is claimed as a dependent by the grandmother for all years of our sample, we would never identify the child. In addition to these two forms of under-counting, we also over-count the number of children because our procedure could miscategorize other dependents as children. Because most such dependents tend to be elderly parents, the fraction of cases that are incorrectly categorized as teenage births is likely to be small. Despite these measurement problems, we believe that the teenage birth variable is reasonably accurate because the aggregate statistics match national averages. Moreover, teenage birth correlates with other observables as expected. For instance, women who score higher on tests, attend college, or have higher income parents are significantly less likely to have teenage births.

Parent Characteristics—We link students to their parents by finding the earliest 1040 form from 1996–2010 on which the student was claimed as a dependent. We identify parents for 88.5% of students linked with tax records as adults. The remaining students are likely to have parents who were not required to file tax returns in the early years of the sample when they could have claimed their child as a dependent, making it impossible to link the children to their parents. Note that this definition of parents is based on who claims the child as a dependent, and thus may not reflect the biological parent of the child.

For the analysis of tax credit receipt on scores, we use contemporaneous parental characteristics to measure EITC and CTC eligibility. Since standardized testing occurs in either January or April of each calendar year (depending on the year and grade), we match a student's test score in year t to the household claiming the child as a dependent in tax year $t - 1$. Therefore, since our tax data begin with tax year 1996, the first year of school data we can use is 1997. If no household claims a student in a given year, we impute forward the last household who claimed the student.[9]

Having matched students to their parents, we can now define a number of key household variables, the most important being household income. If the household has filed a tax return in a given year, we use adjusted gross income (AGI) as our measure of income. We choose this measure since it is the concept of income most closely related to that which determines eligibility for the key credits in our study.[10] For these households, we can also determine the actual credit paid for the EITC and CTC. For the CTC, we include both the "regular" Child Tax credit payments, which are non-refundable, as well as the "additional" payments reflecting the part of the credit that later becomes refundable. For those households that were not required to file a tax return in a given year, we impute household income from W–2s. We also impute EITC and CTC credits based on the imputed household income measure and the number of dependents claimed in the most recent year in which the household filed, correcting for changes in the ages of dependents.[11]

For the analysis of earnings on scores, we define parental household income as average adjusted gross income (capped at $117,000, the 95th percentile in our sample) over the 3 years when the children were 19–21 years old. For years in which parents did not file, we impute parental household income from wages and unemployment benefits, each of which are reported on third-party information forms. We define marital status, home ownership, and 401(k) saving as indicators for whether the parent who claims the child ever files a joint tax return, has a mortgage interest payment, or makes a 401(k) contribution over the period for which relevant data are available. We define mother's age at child's birth using data from Social Security Administration records on birth dates for parents and children. For single parents, we define the mother's age at child's birth using the age of the filer who claimed the child, who is typically the mother but is sometimes the father or another relative.[12] When a child cannot be matched to a parent, we define all parental characteristics as zero, and we always include a dummy for missing parents in regressions that include parent characteristics.

II.C. Analysis Data Structure

The school district and tax records were linked using an algorithm based on standard identifiers (date of birth, state of birth, gender, and names) that is described in the Appendix. Students born after 1990 are too young for us to obtain data on adult outcomes and are excluded from this part of our analysis, though we use these younger cohorts of students to estimate teacher quality.

The linked analysis dataset has one row per student per subject (math or English) per school year. It contains 5.98 million student-year-subject observations and roughly 5.31 million test scores. 4.80 million observations have teacher links. There are 974,686 unique students which appear on average 6.14 times in the data. 89.2% of the observations in the school district data are matched to the tax data. The match rate is uncorrelated with teacher assignment, suggesting that the small degree of attrition is unlikely to produce significant bias.

Each observation in the analysis dataset lists the student's test score in the relevant subject test, demographic information, and class and teacher assignment if available. Each row also lists all the students' available adult outcomes (e.g., college attendance and earnings at each age) as well as parent characteristics. We organize the data in this format so that each row contains information on a treatment by a single teacher conditional on pre-determined characteristics. We account for the fact that each student appears multiple times in the dataset by clustering standard errors as described below.

II.D. Summary Statistics

Table 1 reports summary statistics for the analysis dataset. Table 1 has three parts. The first panel includes variables for students from the school district data; the second panel shows adult outcome variables for students from the tax data; the third panel includes household (i.e., parent characteristics) characteristics from the tax data. Note that these statistics are student-schoolyear-subject means and thus weight students who are in the district for a longer period of time more heavily, as does our empirical analysis. The mean age at which students are observed is 11.7. The mean test score in the sample is positive and has a standard deviation below 1 because we normalize the test scores in the full population that includes students in special education classrooms and schools (who typically have low test scores). Within our analysis sample, 3% of students receive special education services, while 10% have limited English proficiency. Roughly 80% of students are eligible for free or reduced price lunches; 3% of the observations are for students who are repeating the current grade.

TABLE 1: Summary Statistics

Variable	Mean	S.D.	Observations
	(1)	(2)	(3)
Student Data:			
Class size (not student-weighted)	28.3	5.8	211,371
Teacher experience (years)	8.08	7.72	4,795,857
Test score (SD)	0.12	0.91	5,312,179
Female	50.3%	50.0%	5,336,267
Age (years)	11.7	1.6	5,976,747
Free lunch eligible (1999–2009)	76.0%	42.7%	2,660,384
Minority (Black or Hispanic)	71.8%	45.0%	5,970,909
English language learner	10.3%	30.4%	5,813,404
Special education	3.4%	18.1%	5,813,404
Repeating grade	2.7%	16.1%	5,680,954
Number of subject-school years per student	6.14	3.16	974,686
Student match rate to adult outcomes	89.2%	31.0%	5,982,136
Student match rate to parent characteristics	94.6%	22.5%	5,329,715
Adult Outcomes:			
Annual wage earnings at age 20	4,796	6,544	5,255,599
Annual wage earnings at age 25	15,797	18,478	2,282,219
Annual wage earnings at age 28	20,327	23,782	851,451
In college at age 20	36.2%	48.1%	4,605,492
In college at age 25	17.3%	37.8%	1,764,179
College Quality at age 20	24,424	12,834	4,605,492
Contribute to a 401(k) at age 25	14.8%	35.5%	2,282,219
ZIP code % college graduates at age 25	13.2%	7.1%	1,919,115
Had a child while a teenager (for women)	8.4%	27.8%	2,682,644
Parent Characteristics:			
Household income (child age 19–21)	35,476	31,080	4,396,239
Ever owned a house (child age 19–21)	32.5%	46.8%	4,396,239
Contributed to a 401k (child age 19–21)	25.1%	43.3%	4,396,239
Ever married (child age 19–21)	42.1%	49.4%	4,396,239
Age at child birth	27.6	7.4	4,917,740
Predicted Score	0.16	0.26	4,669,069

Notes: Adult outcomes and parent characteristics are from 1996–2010 tax data; student data is from the administrative databases of a large U.S. school district. All ages refer to the age of an individual as of December 31 within a given year. Earnings are average individual wage earnings reported on W–2 forms; those with no W–2 earnings are coded as zeros. College attendance is measured by the receipt of a 1098–T form, issued by a higher education institution to report tuition payments for scholarships. For a given college, "college quality" is defined as the average wage earnings at age 30 in 2009 for the subset of the entire U.S. population enrolled in that college at age 20 in 1999. Individuals who do not attend college are coded as the mean earnings at age 30 in 2009 of all individuals in the U.S. population not in college at age 20 in 1999. Teenage births are measured only for females, by the claiming of a dependent, at any time in our sample, who is fewer than 20 years younger than the individual. Home ownership is measured by those who report mortgage interest payments on a 1040 or 1098 tax form. Marital status is measured by whether an individual files a joint return. Zipcode of residence is taken from either the address reported on 1040 or W–2 forms; for individuals without either in a given year, we impute location forward from the most recent non-missing observation. Percent college graduate in the neighborhood is based on data from the 2000 Census. We link students to their parents by finding the earliest 1040 form from 1998–2010 on which the student is claimed as a dependent. We are unable to link 32% of students to their parents; the summary statistics for parents exclude these observations. Parent income is average adjusted gross income during the tax-years when a student is aged 19–21. For parents who do not file, household income is defined as zero. 401(k) contributions are reported on W–2 forms. Other parent variables are defined in the same way as student variables. All monetary values are expressed in real 2010 dollars. Student data pools grades 4–8, excluding only individuals who are in special education schools or special education classrooms (defined as > 25% special ed. within a classroom).

The availability of data on adult outcomes naturally varies across cohorts. For instance, there are 1.4 million student-subject-school year observations for which we see both teacher assignment and earnings at age 25, about 376,000 at age 28, and only 63,000 at age 30. Hence, age 28 turns out to be the oldest age at which the sample is large enough to obtain precise estimates of teachers' impacts on earnings. Because test score data are available in the spring of 1989 and 1990 but class and teacher assignment data are available only starting in 1991, we are able to examine the cross-sectional relationship between scores and earnings up to age 30.

Mean earnings at age 28 is $20,327 (in 2010 dollars), which includes zero earnings for 34% of the sample. 36% of students are enrolled in college at age 20.[13] Among colleges attended by students in our sample, the average value of our earnings-based index of college quality is $38,623.

For students whom we are able to link to parents, mothers are 28 years old on average when the student was born. One quarter of parents made a 401(k) contribution and nearly one-third own a house at some point when their child was between the ages of 19 and 21. Mean parent household income is $35,476 (in 2010 dollars). Though our sample includes more low income households than would a nationally representative sample, our data include a substantial number of higher income households, allowing us to analyze the impacts of teachers across a broad range of the income distribution. The standard deviation of parent income is approximately $31,080, with 10% of parents earning more than $83,669.

As a benchmark for evaluating the magnitude of the causal effects estimated below, Appendix Tables 3–6 report estimates of ordinary least squares (OLS) regressions of the adult outcomes we study on test scores. Both math and reading test scores are highly positively correlated with earnings, college attendance, and neighborhood quality and are negatively correlated with teenage births. In the cross-section, a 1 SD increase in test score is associated with a $7,440 (37%) increase in earnings at age 28. Conditional on lagged test scores and other controls that we use in our empirical analysis below, a 1 SD increase in test score is associated with $2,545 (11.6%) increase in earnings. We show below that the causal impact of teacher value added (VA) on earnings is similar to what one would predict based on the cross-sectional correlation between scores and earnings conditional on controls.

III. Estimates of the Impact of Tax Credits on Student Achievement

We identify the effect of tax-related transfers on educational achievement by exploiting the non-linearities in the schedule of key credits. We first describe the two key policies from which we derive identification. We then specify our estimating equations and present graphical results, followed by regression estimates.

III.A. Earned Income Tax Credit

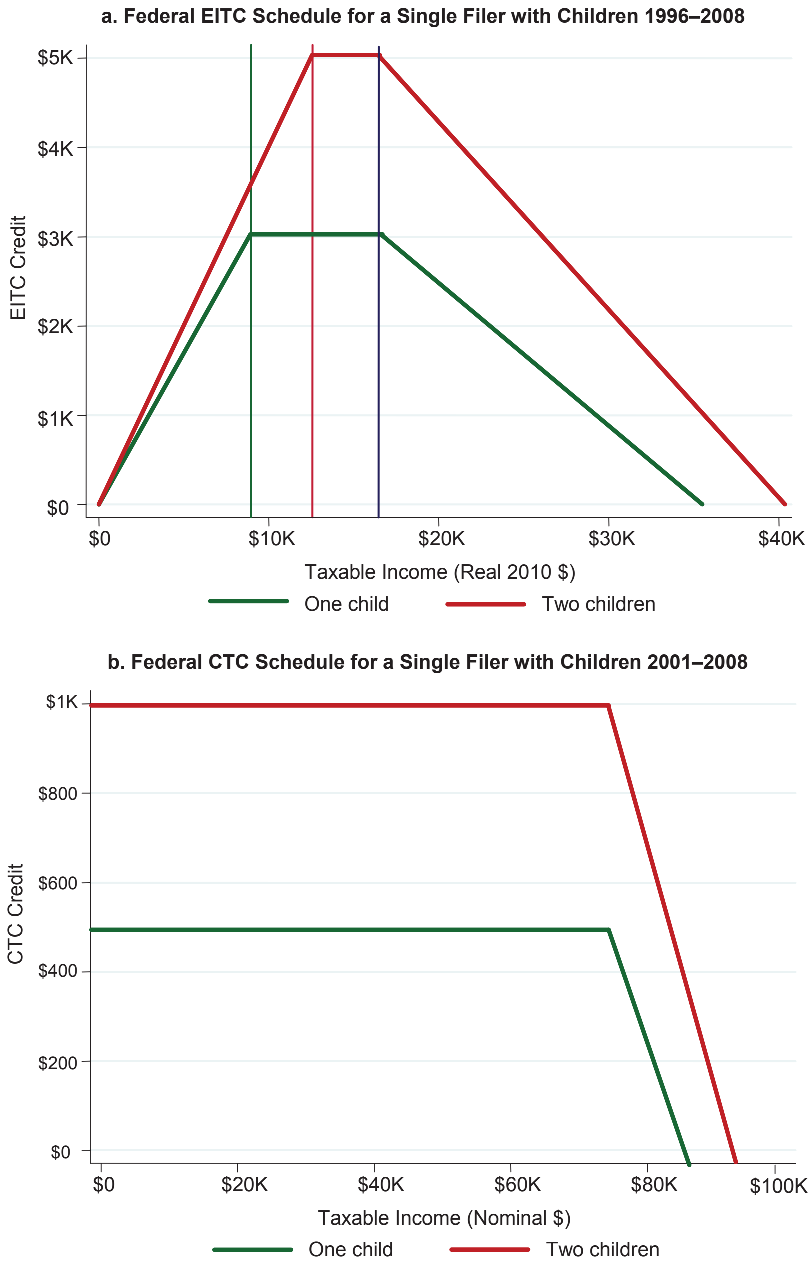
The Earned Income Tax Credit (EITC) is a refundable credit paid to households with positive income. The defining feature of the EITC is the pyramid-shaped schedule of the credit, displayed in Figure 1a.[14] Households receive an income subsidy for earnings up to a certain threshold. For all years in our sample, households with one dependent receive a 34% credit up to $8,970 of income for a maximum credit of $3,050.[15] Household with two children or more receive a 40% credit up to $12,590 for a maximum credit of $5,036.[16] Eligible dependents must live in the household for more than 6 months during a given tax year, and must remain either under 19 or full-time students under 24 for the entire tax year. The credit is then phased out once households earn more than $16,690; the phase-out rate is 16% for households with one child, and 21% for households with two or more children. Beginning in 2002, Congress lengthened the "plateau" range, while leaving the phase-out rate unchanged. In the first three years after the reform, the phase-out begins for married households at $17,690; in 2005–2007, the phase-out period begins at $18,690; and in 2008, the phase-out begins at $19,690.

Table 1 provides summary statistics on the tax credit parameters. Families in our sample earned an average of $1,520 from the EITC. Note that this figure excludes households that do not qualify for the EITC. Approximately 65% of families in our sample qualify for the credit.

III.B. Child Tax Credit

For families with earnings below an income threshold, the Child Tax Credit (CTC) provides a partially refundable credit for each eligible dependent. Figure 1b depicts the credit schedule for a single filer for 2001–2008. The size of the basic credit is constant below the income threshold; after passing the threshold, the phase-out rate is 5%. The income threshold is $75,000 for singles and $110,000 for married households filing jointly. The

FIGURE 1: EITC and CTC Schedules



Notes: Panel a. displays the Earned Income Tax Credit schedule for 1996–2008 for households filing as head of household with children, in constant 2010 dollars. Panel b. displays the maximum Child Tax Credit for which a household is eligible from 2001–2010. In both panels, the green line shows the schedule for those with one dependent; the red line shows the schedule for those with two dependents in nominal dollars.

CTC offered $400 per child (up to two) in 1998 (the first year of the credit), $500 in 1999–2001, and $1,000 from 2002 on, in nominal dollars.

Before 2001, the CTC was non-refundable. Since many low-income families owe no income tax, they could not benefit from the CTC. Beginning in 2001, the CTC became partially refundable, where the newly refundable portion was called the Additional Child Tax Credit. Households were able to claim a refundable credit up to 15% of their income above an income threshold of $12,050. For example, consider a family with two children and $22,050 of taxable income that owed $300 in tax payments (after the EITC). Under the original CTC, this family would claim \$300 in offsetting credit but could not claim more. Under the Additional CTC, the family could claim an additional amount equal to 0.15 × (\$22,050 − \$12,050) = \$1,500. The family would thus receive a CTC equal to \$1,800 in total.

Given the many potentially endogenous factors that enter into the deductibility of the CTC, it is not appropriate to use the actual CTC payment for identification. For instance, parents might claim another credit that would limit refundability but not reduce the overall payments. Therefore, we use the simulated credit—that is, the maximum credit possibly due—for identification. Since the simulated CTC is constant for much of the range of our analysis, the use of the simulated CTC places much of the burden for identification on the EITC in practice.

Table 1 provides summary statistics on the tax credit parameters. On average in our sample, families qualify for \$606 from the Child Tax Credit, which is the non-refundable portion of the credit, plus an additional \$537 from the Additional Child Tax Credit. Approximately 82% of families qualified for either the Child Tax Credit or the Additional Child Tax Credit. Combining both credits, we find that families received an average of \$1,652 from the EITC and CTC combined. This represents approximately 60% of total tax credits.

III.C. Estimating Equation and Identification Assumptions

Both the EITC and CTC have highly non-linear schedules. In contrast, other determinants of a child's achievement change more smoothly throughout the income distribution. Our basic estimating equation is therefore

$$A_{ift} = \alpha + \phi\,(AGI_{ft}) + \beta * CREDIT_{ft} + \gamma X_{ift} + \varepsilon_{ift} \qquad (1)$$

for student *i* in family *f* in year *t*, where A_{ift} is achievement on the standardized test at the end of the year, $\phi\,(\cdot)$ is a smooth function of family *AGI*, *CREDIT* is the combined EITC and simulated CTC payments to family *f* in year *t*, and X_{ift} is a vector of individual and family characteristics including English proficiency, receipt of special education status, age, and gender, as well as the household background characteristics including a dummy variable for married filing status, a dummy variable for the difference between the age of the claiming parent and dependent less than 20 years, a dummy variable for home ownership, and average savings in tax-deferred account. In practice, we use a five-degree polynomial to estimate the smooth function $\phi\,(\cdot)$. We have also run similar specifications using higher-order polynomials, as well as smoothed splines (i.e., splines that have continuous derivatives at knot-points), and the results are unaffected.

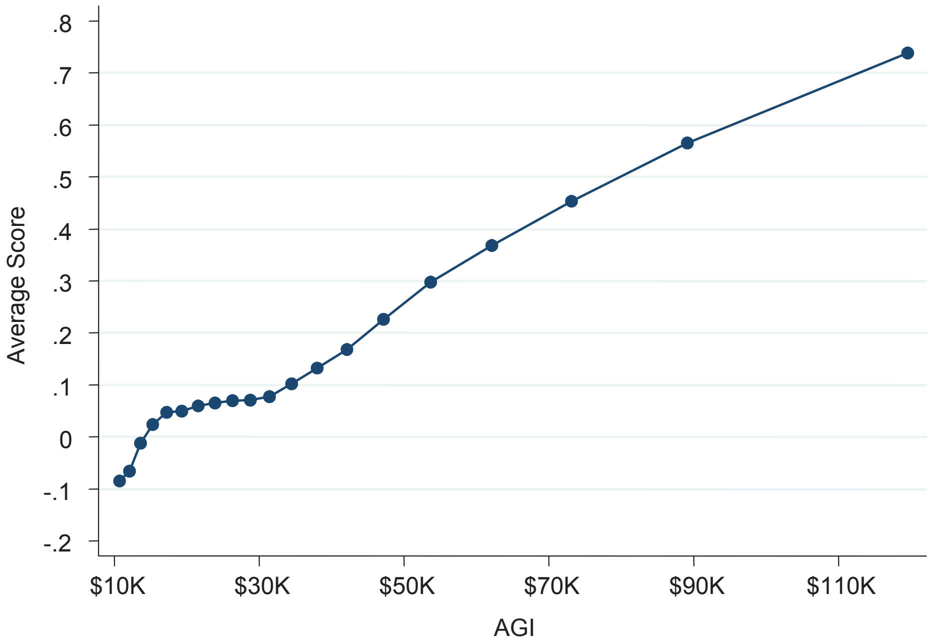
Our key identification assumption is that the smooth function $\phi\,(\cdot)$ captures the entire relationship between simultaneous parental income and achievement other than that driven through the receipt of federal credits. In practice, the EITC provides most of the identification in our study. The key identification question may therefore be restated as: Do children of families earning between roughly \$10,000 and \$30,000 in AGI overperform in school, relative to the trend determined by their higher and lower scoring peers? Although it is difficult to rule out all confounding effects, the analysis below suggests that the over-performance of children in the EITC range hues quite close to the actual schedule and is therefore unlikely to be generated by other phenomena. Nevertheless, an important question for future research is to affirm these results. Unfortunately, the available tax data exist only back to 1996, just after the large EITC reforms of the mid-1990s. It is therefore impossible in this study to rely on the sharp changes in the EITC credit (as in Dahl and Lochner, 2011).

III.D. Graphical Evidence

We begin by plotting the cross-sectional patterns of the two key variables: household income and student achievement. Figure 2 plots average scores, as a function of contemporaneous household income. Overall, scores are sharply increasing with household income, with an average slope of approximately 0.01. This implies that each \$10,000 of income increases scores by roughly 0.1 SD. The relationship between scores and income is generally quite smooth and slightly concave, except between \$10,000 and \$30,000.

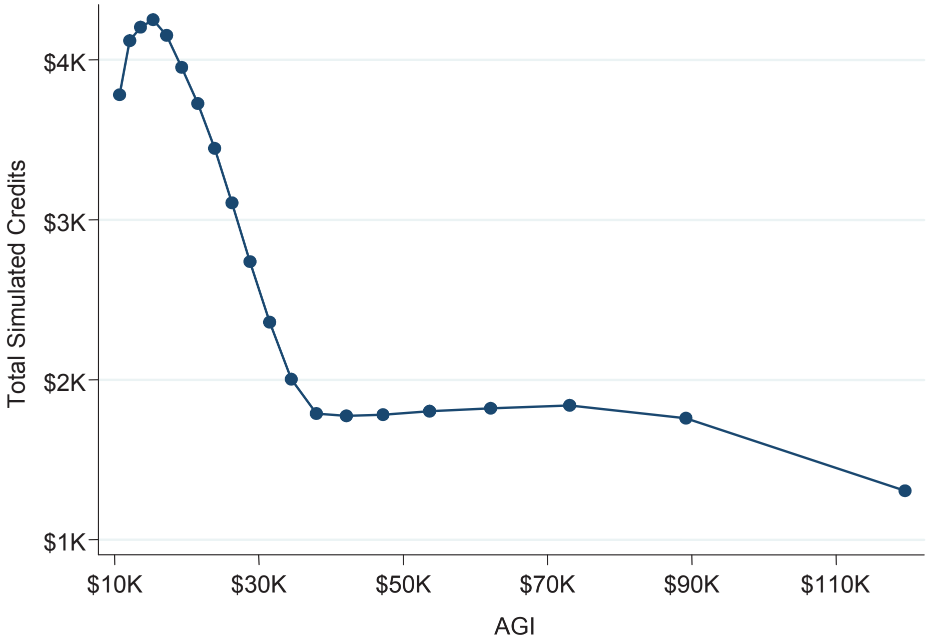
Figure 3 presents simulated credits as a function of AGI. The shape of the EITC is clear at the lower end of the income distribution, where the simulated credit first rises and then sharply falls as a function of AGI.

FIGURE 2: Average Score vs. AGI



Notes: This figure plots the average test scores as a function of household adjusted gross income (AGI). Test scores come from a large school district between 1997 and 2009. We normalize test scores within grade-year cells and then average reading and math scores for each student within a year. Household adjusted gross income comes from 1040 forms accessed through selected federal tax records. See Appendix A for details on the procedure by which we match students to households in the tax data. To construct this figure, we group student-year observations into 20 equal-sized (5 percentile point) bins and plot the mean test score for each bin. All monetary values are expressed in real 2010 dollars.

FIGURE 3: Total Simulated Credits vs. AGI



Notes: This figure plots the average simulated credit amount from the Earned Income Tax Credit and the Child Tax Credit by household adjusted gross income (AGI). To construct this figure, we group student-year observations into 20 equal-sized (5 percentile point) bins and plot the mean test score for each bin. All monetary values are expressed in real 2010 dollars.

Above about $40,000, the simulated credit flattens, reflecting the constant credit available in this income range through the CTC. The credit then falls gradually again once incomes begin to rise above the threshold for the CTC. In contrast with the sharp kinks present in the credit schedules in Figure 1, note that the simulated credit amounts in Figure 3 show a similar but smoothed pattern. This is because Figure 3 averages over single and married households, as well as households with different numbers of children. This serves to smooth the credit schedules slightly.

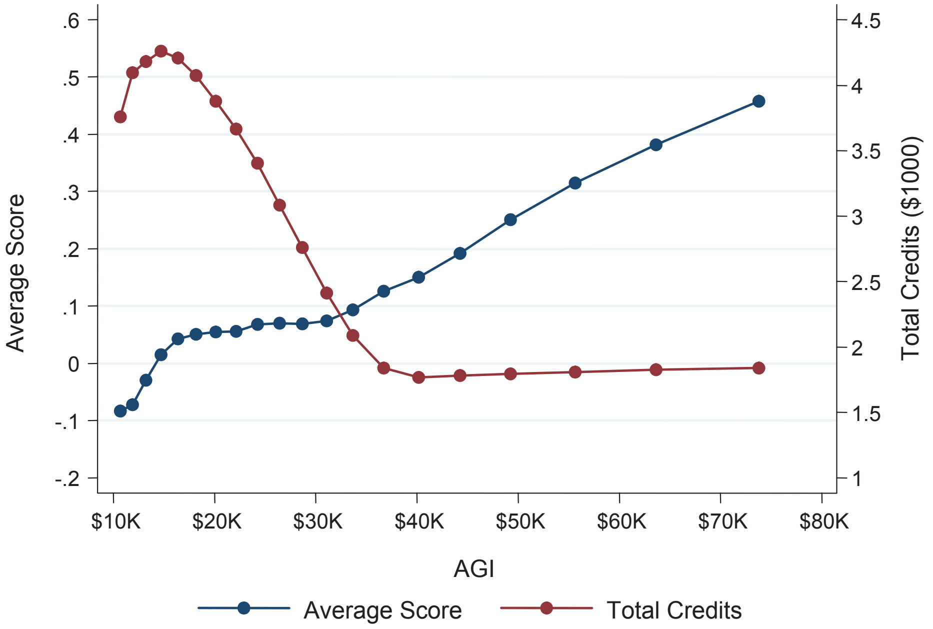
Figure 3 also makes clear that the main identification in this paper comes from the EITC, rather than the CTC. The EITC appears as a dramatic increase and decrease of available credit, while the CTC appears as a slight decline in the credit. The reason for this difference is apparent in the phase-in and phase-out rates present in each program. The CTC is a constant credit with a 5% phase-out rate at the end. In contrast, the EITC provides phase-in and phase-out rates that are several times higher. As a result, the marginal effects of the CTC are simply too small to be noticeable on the necessary scale.

Figure 4 combines Figures 2 and 3 and zooms in on the lower end of the income range where there is the most variation in the credit to permit direct comparison of the cross-sectional patterns. Above $40,000, both series are smooth and roughly linear. Below there, however, each series shows a striking deviation from the otherwise smooth pattern. It is in this range that the EITC more than doubles the credit available to households with children. And it is also in this range that children appear to over-perform significantly relative to the income-achievement gradient established in the rest of the figure. Furthermore, the break in linearity in each figure occurs at exactly the same place. Just as the simulated credit available through the EITC begins to increase, student achievement turns up from projected path. We now proceed to explore the relationship more formally.

III.E. Identification Approach 1: Cross-Sectional Identification from Policy Non-Linearities

We estimate equation (1) in Table 2. Column 1 estimates the most parsimonious specification, including only a linear control for household AGI. The coefficient is 0.075 SD, and is highly significant with a standard error of 0.002, implying a t-stat of approximately 37. Column 2 increases the flexibility of the AGI control function. Now using a cubic function of AGI as the function ϕ (·) in equation (1) we estimate exactly the same coefficient and standard error.[17]

FIGURE 4: Average Score and Total Credits vs. AGI



Notes: This figure plots the average test score and credit amount for each bin of household AGI. See Figures 2 and 3 for details on the construction of these variables. To construct this figure, we group student-year observations into 20 equal-sized (5 percentile point) bins and plot the mean test score for each bin. All monetary values are expressed in real 2010 dollars.

TABLE 2: Impacts of Tax Credit on Test Scores

	Test Score				
Dependent Variable:	Linear AGI	Cubic AGI	Full Controls	Math	Reading
	(1)	(2)	(3)	(4)	(5)
Simulated Credits	0.075	0.075	0.080	0.093	0.062
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	3,006,098	3,006,098	3,006,098	1,533,339	1,472,759

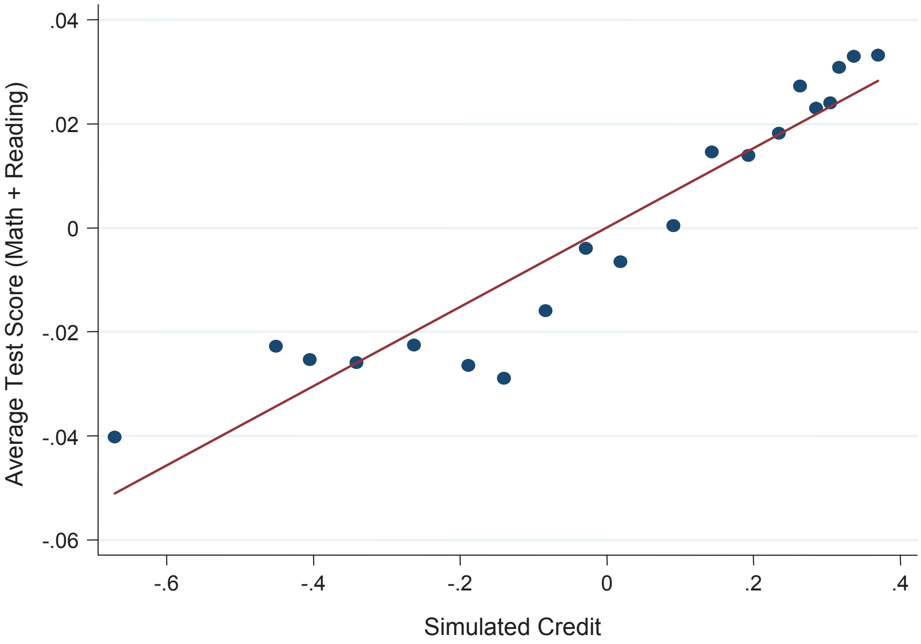
Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. In all columns, the dependent variable is student test score, normalized within grade-year cell. Columns 1–3 include the entire sample and differ only in the control variables included. Column 1 includes only a linear control for household AGI. Column 2 includes only a quadratic polynomial control for household AGI. Column 3 includes a five-degree polynomial in household AGI, as well as all school and household control characteristics. Column 4 replicates the regression in Column 3 using only math scores. Column 5 replicates the regression in Column 3 using only reading scores.

Intuitively, the best-fit curve for the relationship between income and score is highly linear, even when the regression allows for more flexibility. Therefore the linear and cubic specifications yield nearly identical results.

Column 3 presents our primary specification. In it we control for a quintic polynomial of AGI, as well as the vector of individual characteristics described above. These additional controls increase the coefficient of interest slightly to 0.08 SD. The coefficient increases because individuals who receive lower amounts of credits not only have more household income, but also tend to be from households with married parents and mothers who gave birth at a later age. Each of these additional characteristics also predicts higher test scores; when controlling for them, tax credits appear to have an even larger impact on achievement.

Figure 5 represents the regression in Column 3 of Table 2 in scatterplot form. We regress both achievement and simulated credit on the polynomial in AGI and other controls, and then take residuals. We then

FIGURE 5: Average Test Score vs. Credit



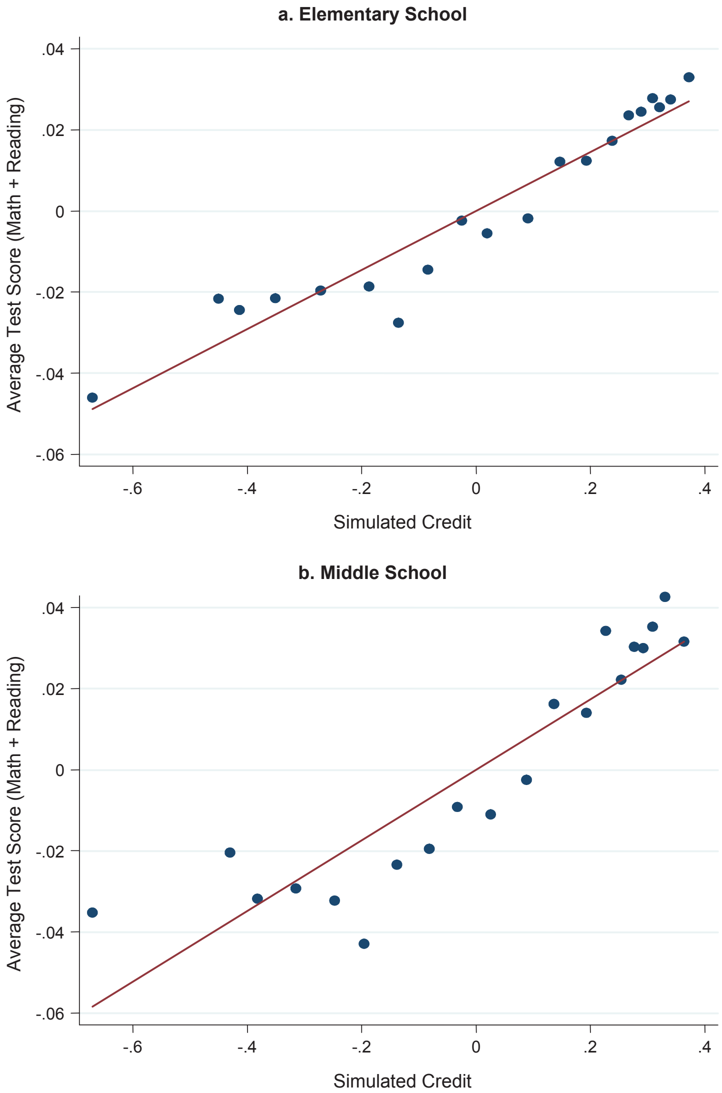
Notes: This figure represents the data underlying the regression in Column 3 of Table 2. We regress both average test score and household credit amount on the full vector of control variables, including a five-degree polynomial in household income, student English proficiency, receipt of special education status, age, and gender, as well as the household background characteristics including a dummy variable for married filing status, a dummy variable for the difference between the age of the claiming parent and dependent less than 20 years, a dummy variable for home ownership, and average savings in tax-deferred account. We add back the sample means of each variable, and then group observations into 20 equal-sized (5 percentile point) bins and plot the mean residual test score in each bin. The best-fit line is that fitted on the underlying individual data.

group observations into 20 bins based on the size of the tax credit residual and plot the mean achievement for students in each bin. Intuitively, Figure 5 presents a non-parametric version of the key regression coefficient in Column 3. The linear fit appears approximately correct, and the relationship is not driven by outliers in either direction.

Our estimated coefficient of 0.08 is large when compared with the cross-sectional impact of income. Income transfers appear more than an order of magnitude more effective in increasing student test scores. Dahl and Lochner (2011) point out that a $1,000 increase in the EITC, for instance, is likely to be a far more permanent income shock than a $1,000 increase in earned income. Furthermore, the income elasticity of test scores is likely to be far higher at the lower reaches of the income distribution, which one can partly see from the concavity of the underlying relationship between scores and income in Figure 2.

It is also worth noting at this time the strong assumption on the cross-sectional pattern of test scores and household income on which our identification strategy depends. This relationship must hold constant across

FIGURE 6: Average Test Score vs. Credit by School



Notes: This figure replicates Figure 5, after splitting the observations from Figure 5 into those from grades 3–5 (Panel a.) and grade 6–8 (Panel b.). See the notes to Figure 5 for details.

high and low-income households in order for our identification method to be valid. We discuss below a number of specific ways in which this assumption may be violated, and thus these results should be interpreted with caution.

Columns 4 and 5 repeat the specification in Column 3, separating out math and reading tests. The results suggest that income from tax credits has a larger impact on math scores than reading scores. Each $1,000 in income generates a 0.093 SD increase in math scores, but only a 0.062 SD increase in reading scores. There are a number of possible explanations for this finding. First, previous research has documented that math scores are more highly correlated with long-term outcomes than reading scores. Under this interpretation, math scores are the best proxy for underlying academic achievement, while reading scores reflect more a child's family background or worldly exposure than true ability. Second, it is possible that reading scores are a particularly bad proxy for achievement in a low-income population where English is often a second language.

Table 3 investigates heterogeneity of these effects across grades. Each column replicates the specification in Column 3 of Table 2, restricting to a single grade. We find that the impact of federal credits increases in later grades, though the effect is non-monotonic. The effect size starts at 0.073 in 3rd grade, after which it actually falls to 0.069 in 4th grade. The effect then begins to rise, and we measure the effects in 5th grade to be 0.078. We do not have a theory to explain this variation; however, neither the 4th nor 5th grade coefficients are statistically significantly different from the 3rd grade effect, suggesting an average effect of 0.073 for all of elementary school.

TABLE 3: Heterogeneity by Grade

Dependent Variable:	Test Score					
	(1)	(2)	(3)	(4)	(5)	(6)
Grade	3	4	5	6	7	8
Simulated Credits	0.073	0.069	0.078	0.082	0.085	0.090
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)
Observations	638,903	623,058	555,894	476,987	396,939	314,280

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. The columns replicate Column 3 from Table 2 for each grade separately. See the notes to Table 2 for details.

We run the same specifications for middle school grades in the next three columns, which tell a different story. Here the effect of federal tax credits increases consistently from 0.082 in 6th grade to 0.085 in 7th grade and 0.090 in 8th grade. This suggests that as students age, they are better able to take advantage of the benefits afforded through income transfers. Figure 6 graphically displays these two regressions, confirming that the effect of tax credits on scores is larger in middle school

Having estimated the basic effect of tax credits on achievement, we now turn to time-series variation in the size of credits to confirm our estimates.

III.F. Identification Approach 2: Time-Series Variation from Policy Reform

Our second identification strategy exploits increases over time in the generosity of state and local matches to the federal EITC. Conceptually, there are two ways to examine the effect of such increases: panel analysis and repeated cross-sectional analysis. To understand the intuitive difference between these two methods, consider a family who is EITC-eligible in the year before the reform. In order to measure the impact of the policy change, we must compare the achievement of the pre-reform family to someone before the reform. The panel analysis approach seeks to compare our pre-reform family to the same family, post-reform. If incomes were relatively stable, this method would be an attractive way of controlling for unobservable qualities of a family environment. But incomes, especially at the low end of the distribution, often vary wildly across the years. As a result, the panel analysis method must control for large changes in family income while examining the impact of the reform. In contrast, the repeated cross-sectional analysis seeks to compare our pre-reform family to a different family who is at the same income level after the reform. This approach must rely more heavily on controls to account for differences in family structure since it cannot include family fixed effects. The advantage is that one need not attempt to model the complex process of mean-reversion in income. A similar methodological debate has recently occurred in the literature measuring the tax elasticity of labor supply; the most recent papers in this literature have concluded that the cross-sectional approach is more robust (Saez et al., 2012). We therefore adopt a repeated cross-sectional approach here.

In the state we studied, there were a number of policy changes through our period that increased the effective size of the federal credits. These changes occurred somewhat continuously up until 2006, when the largest increase occurred. Policy was relatively stable after 2006. Therefore, we should expect the impact of federal credits to increase up through 2006, after which it should remain stable.

Table 4 presents the results of our time-series analysis. We find that the impact of federal tax credits on achievement increases sharply from 2003, when the effect estimate is only 0.037, through 2006 when we estimate the effect at 0.097. The coefficients then level off through 2007 and 2008. Figure 7 depicts the change in estimates over time. The effect of federal tax credits clearly increases up through the policy change in 2006, but levels off thereafter. Interestingly, the increases in the coefficient from 2003 to 2006 seems "too large" relative to the average effect that we estimated in Table 2. However, it is possible that the state and local reforms in those years better targeted families who could use the money best to increase student achievement. It is also possible that these state and local income transfers were accompanied by other policies designed to support low income children at school, or perhaps even targeted changes in the school system itself. Therefore we consider these effects qualitatively consistent with our earlier findings, though quantitatively inconclusive on the impact of the state and local match programs.

TABLE 4: Changes in Impacts of Credits Over Time

Dependent Variable:	Test Score					
	(1)	(2)	(3)	(4)	(5)	(6)
Year	2003	2004	2005	2006	2007	2008
Simulated Credits	0.037	0.051	0.066	0.097	0.100	0.104
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)
Observations	249,290	330,098	397,792	461,961	480,449	467,816

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. The columns replicate Column 3 from Table 2 for each tax year separately. See the notes to Table 2 for details.

IV. Estimates of the Impact of Score Gains on Adult Outcomes

We now turn to the link between scores and earnings. We need variation in scores caused by a randomized policy intervention in order to estimate the causal impact of scores on earnings. We use teacher assignment to generate such variation because the students who are affected by the tax credits themselves in our sample are too young to have earnings data.

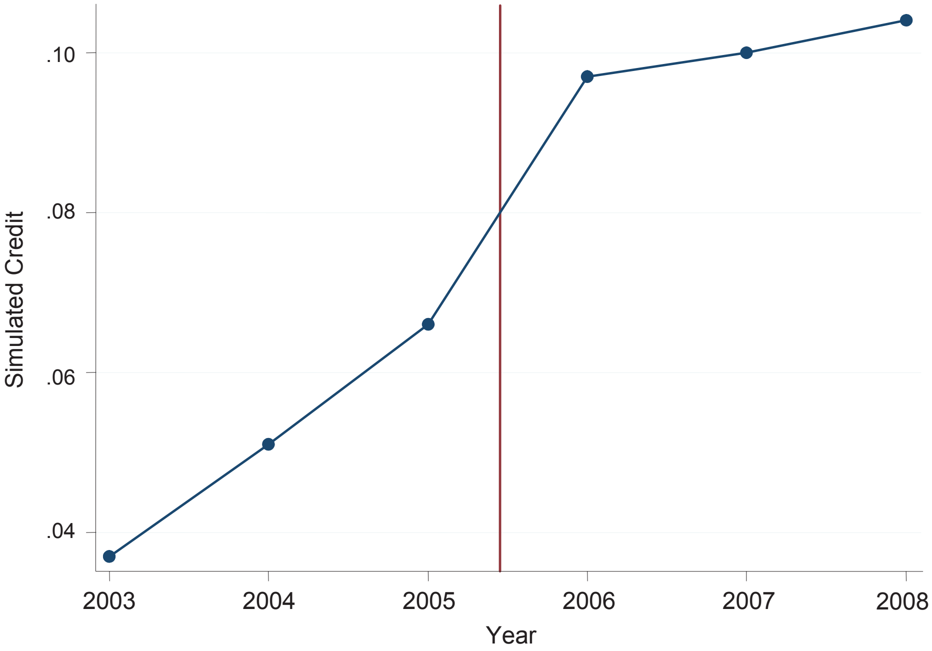
We first measure the impact of each individual teacher on scores. In particular, for a given classroom we use a teacher's effect on scores in other classrooms she teaches to measure the effect on scores. To do so, we estimate the following empirical model of test scores:

$$A_{igt} = f_{1g}(A_{i,t-1}) + f_2(\bar{A}_{c(i,g),t-1}) + \phi_1 X_{igt} + \phi_2 \bar{X}_{c(i,g)} + \nu_{igt}. \tag{2}$$

Here $f_{1g}(A_{i,t-1})$ denotes a control function for individual test scores in year $t-1$, $f_2(\bar{A}_{c(i,g),t-1})$ denotes a control function for mean classroom test scores in year $t-1$, X_{igt} is a vector of observed individual-level characteristics (such as whether the student is a native English speaker), and $\bar{X}_{c(i,g)}$ is a vector of classroom-level characteristics determined before teacher assignment (such as class size or a dummy for being an honors class). For classrooms in a given year t, we then measure teacher quality as the average residual in other years so that

$$\hat{\mu}_{jt} = E\,[\nu_{igt'} \mid t \neq t'\,]\,.$$

FIGURE 7: Changes in Impacts of Credits Over Time



Notes: This figure plots six regression coefficients. For each coefficient, we run the regression detailed in Column 3 of Table 2 on the sample from each tax year. See notes to Table 2 for details.

After calculating this average residual, we then scale the teacher quality measures so that a rating of 0.1 implies that a teacher is predicted to increase student scores by 0.1. This corrects for measurement error. For a more detailed description of this procedure, see Kane and Staiger (2008).

We can then run regressions of the form

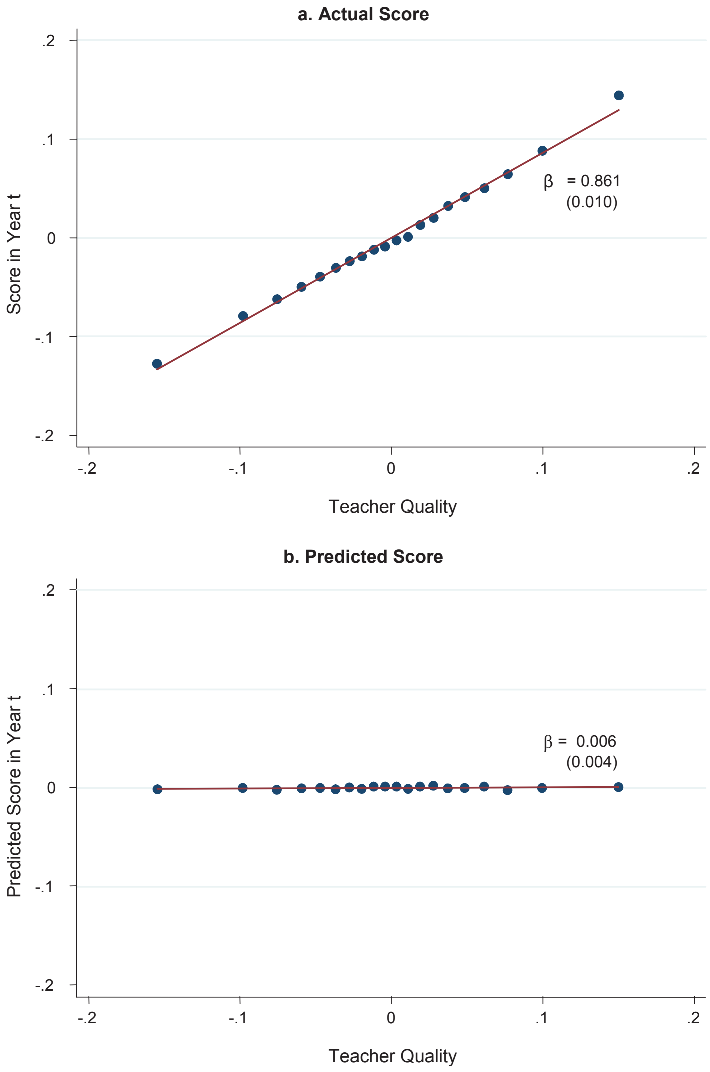
$$Y_i = \beta\hat{\mu}_{j(i,g)} + f^{\mu}_{1g}(A_{i,t-1}) + f^{\mu}_{2}(\bar{A}_{c(i,g),t-1}) + \phi^{\mu}_{1} X_{igt} + \phi^{\mu}_{2} \bar{X}_{c(i,g)} + \varepsilon^{\mu}_{igt}$$

which includes the same set of controls as equation (2) but also includes teacher quality $\hat{\mu}$. The coefficient β can then be interpreted as our parameter of interest: what is the earnings gain predicted from an increase in student test scores?

Our outcomes have a correlated error structure because students within a classroom face common class-level shocks and because our analysis dataset contains repeat observations on students in different grades. One natural way to account for these two sources of correlated errors is to cluster standard errors by both student and classroom (Cameron et al., 2008). Unfortunately, implementing two-way clustering on a dataset with 5 million observations was infeasible because of computational constraints. We instead cluster standard errors at the school-cohort level, which adjusts for correlated errors across classrooms, and repeat student observations within a school. Clustering at the school cohort level is convenient because it again allows us to conduct our analysis on a dataset collapsed to class means. We show in Appendix Table 7 that in smaller subsamples of our data, school-cohort clustering yields slightly more conservative confidence intervals than two-way clustering by class and student. In addition, we show that our main estimates remain statistically significant when clustering by classroom in a sample that includes only the first observation for each student.

Finally, in our baseline specifications, we exclude teachers whose estimated quality falls in the top 2% for their subject (above 0.21 in math and 0.13 in English) because these teachers' impacts on test scores appear suspiciously consistent with cheating. Jacob and Levitt (2003) develop a proxy for cheating that measures the extent to which a teacher generates very large test score gains that are followed by very large test score losses for the same students in the subsequent grade. Jacob and Levitt establish that this is a valid proxy for cheating

FIGURE 8: Effects of Teacher Assignment on Actual and Predicted Scores



Notes: These figures plot student scores against teacher assignment. Panel a. plots contemporaneous student scores. Panel b. plots predicted scores based on parent characteristics. To construct predicted score, we predict values from a regression of score on a quartic in parent mean income (measured during the years when a student is 19–21), a dummy variable for whether the parents are married during the 3 years when a student is 19–21, an interaction of the quartic with the married dummy, and dummy variables for parent home ownership when the student is 19–21, the age difference between the claiming parent being less than 20 years, and parents' contribution to tax-deferred savings accounts when the student is 19–21. All figures adjust for the full vector of control vectors, including a flexible polynomial n lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, teacher experience, and year- and grade-level fixed effects. To construct each plot, we regress both the y- and x-axis variable on the control vector to calculate residuals. We then group the observations into 20 equal-sized (5 percentile-point) bins based on the x-axis residual and plot the average value of both the y- and x-axis residuals within each bin, adding back the sample means of each variable for ease of interpretation. The solid line shows the best linear fit estimated on the underlying data using

using data on unusual answer sequences that suggest test manipulation. Teachers in the top 2% of our estimated VA distribution are much more likely to be cheating as defined by Jacob and Levitt's proxy. We therefore trim the top 2% of outliers in all the specifications reported in the main text. We investigate how trimming at other cutoffs affects our main results in Appendix Table 8.

IV.A. Is Teacher Assignment Correlated with Other Factors?

Teacher assignment provides consistent estimates of the long-run impact of score gains only if unobserved determinants of students' test score gains do not differ systematically across teachers. If some teachers are assigned better-performing students than others, our estimates will incorrectly reward or penalize teachers for the mix of students they get. Recent studies by Kane and Staiger (2008) and Rothstein (2010) among others have reached conflicting conclusions about whether these estimates are biased by student sorting. In this section, we revisit this debate and present new tests for bias in these estimates.

We begin by forecasting test score gains outside the sample used to estimate $\hat{\mu}_j$ to verify that our estimates of teacher quality have predictive power. Under our assumption that true teacher effects μ_j are time-invariant, a 1 SD increase in $\hat{\mu}_j$ should be associated with a 1 SD increase in test scores out of sample. Figure 8a plots student test scores (combining English and math observations) vs. our leave-out mean Empirical Bayes estimate of teacher assignment. Figure 8a shows that a teacher with $\hat{\mu}_j = 1$ in fact generates a 0.861 SD increase in students' test scores out-of-sample. This confirms that our estimates of teacher quality are highly predictive of student test scores. The coefficient on $\hat{\mu}_j$ is slightly below 1, consistent with the findings of Kane and Staiger (2008), most likely because teacher assignment is not in fact a time-invariant characteristic.

1. Testing for Selection

The relationship between $\hat{\mu}_j$ and students' test scores in Figure 8a could reflect either the causal impact of teachers on achievement or persistent differences in student characteristics across teachers. For instance, $\hat{\mu}_j$ may forecast students' test score gains in other years simply because some teachers are systematically assigned students with higher income parents. A natural test for such selection is to examine the correlation between teacher assignment and variables omitted from our model. The parent characteristics from the tax data are ideal to test for selection because they have not been used to measure teacher quality but are strong predictors of student achievement. We collapse the parent characteristics into a single index by regressing test scores on a quartic in parent's household income interacted with an indicator for the filing parent's marital status as well as mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for having no parent matched to the student. Let $\hat{A}_c^p$ denote the class-average predicted test score for class c from this regression. These predicted test scores are an average of the parent characteristics, weighted optimally to reflect their relative importance in predicting test scores.

Figure 8b plots $\hat{A}_{c,g-1}^p$ against $\hat{\mu}_j$, with teacher quality measured using a leave-out mean as described above. Parent characteristics are uncorrelated with teacher assignment conditional on school-district observables $\bar{X}_c$. At the upper bound of the 95% confidence interval, a 1 standard deviation increase in teacher VA raises predicted scores based on parent characteristics by 0.01 SD. This compares with an actual score impact of 0.861 SD. The regression represented in Figure 8b appears in Column 2 of Table 5.

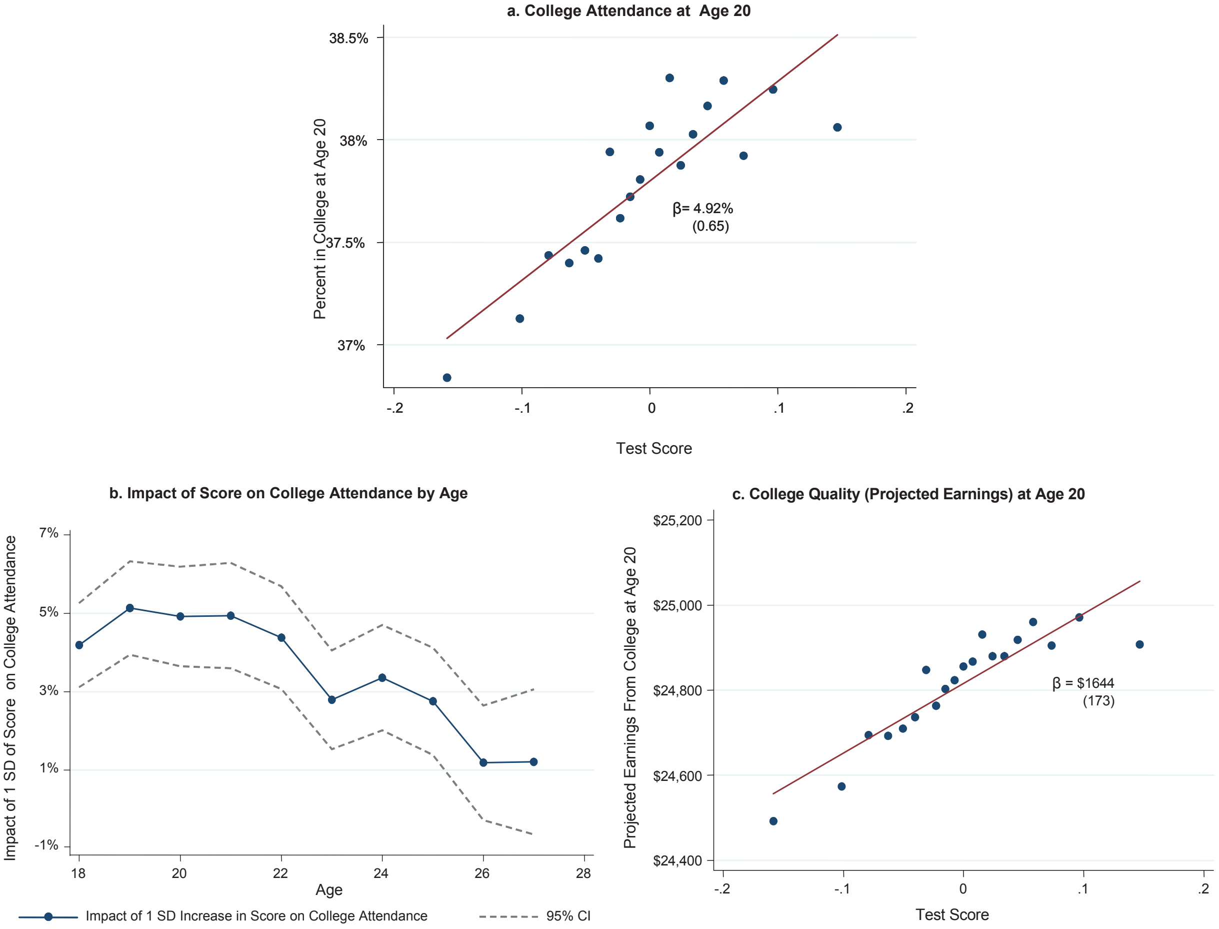
TABLE 5: Tests for Balance Using Parent Characteristics

Dependent Variable:	Score in year t (%)	Predicted Score (%)	Score in year t (%)	Percent Matched (%)
	(1)	(2)	(3)	(4)
Teacher Assignment	0.861	0.006	0.864	0.002
	(0.010)	(0.004)	(0.011)	(0.003)
	[82.68]	[1.49]	[75.85]	[0.562]
Predicted Score Based on Parent Characteristics			0.175	
			(0.012)	
			[62.70]	
Controls	x	x	x	

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. The sample includes all students who would have been in 8th grade by 2004 had they progressed according to the normal pace. There is one observation for each student-year-subject. In Columns 1 and 3, the dependent variable is the student's test score in a given year and subject. In Column 2, the dependent variable is the predicted value generated from a regression of score on a quartic in parent mean income (measured during the years when a student is 19–21), a dummy variable for whether the parents are married at some time during the sample, an interaction of the quartic with the married-dummy, a dummy variable for parent home ownership, a dummy variable for the age difference between the claiming parent being less than 20 years, and a dummy variable for parents' contribution to tax-deferred savings accounts. The second independent variable in Column 3 is the same predicted score from parent characteristics. All columns include the full vector of controls, including a flexible polynomial in lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, teacher experience, and year- and grade-level fixed effects. We cluster standard errors at the school-cohort level.

A closely related method of assessing selection on parent characteristics is to control for class-average predicted scores $\hat{A}_c^p$ when estimating the impact of teachers on actual scores. Columns 3 and 4 of Table 5 shows that the coefficient on $\hat{\mu}_j$ changes from 0.866 to 0.864 after controlling for predicted scores changes when restricting to the sample in which both score and predicted score are present. This robustness to parent characteristics is consistent with the result in Figure 8b. Note that parent characteristics have considerable predictive power for test scores even conditional on $\bar{X}_c$; the t-statistic on the predicted score $\hat{A}_c^p$ is 15. The fact that parent characteristics are strong predictors of residual test scores yet are uncorrelated with $\hat{\mu}_j$ suggests that the degree of bias in VA estimates is likely to be modest (Altonji, Elder, and Taber, 2005).

FIGURE 9: Effect of Test Scores on College and College Quality



Notes: Panel a. plots the relationship between score and the probability her students attend college at age 20, using teacher quality as an instrument. College attendance is measured by receipt of a 1098–T form, issued by higher education institutions to report tuition payments or scholarships, in the year during which a student turned 20. Teacher quality is measured for the teacher of a given class. Panel b. replicates the specification in Panel a., varying the age of college attendance from 18 to 27. Each dot represents the coefficient estimate on score from a separate regression. The dashed lines show the boundaries of the 95% confidence intervals for the effect of value-added at each age. Panel c. plots the effect of score on our earnings-based index of the quality of the college the student attends at age 20. College quality is constructed using the average wage earnings at age 30 in 2009 for all students attending a given college at age 20 in 1999, and denoted in real 2010 dollars. For individuals who did not attend college, we calculate mean wage earnings at age 30 in 2009 for all individuals in the U.S. aged 20 in 1999 who did not attend any college. In Panels a. and c., we calculate residuals for both college attendance and score from a regression on the full control vector, adding back the sample means for ease of interpretation. We then group the observations into 20 equal-sized (5 percentile-point) bins based on the x-axis variable and plot the average value of both the y- and x-axis variables within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. In all panels, standard errors are clustered at the school-cohort level.

Another method of testing the orthogonality of teacher assignment is to include the parental characteristics vector in the control vector when calculating teacher quality. Table 6 shows that the correlation between this new measure of teacher quality and the baseline measure is 0.999. Table 6 also shows that including only lagged scores generates a measure of teacher quality with correlation 0.9 baseline, demonstrating that lagged score is the most important control. In contrast, omitting lagged score generates a measure that is far less correlated with any of the others. We conclude based on these tests that selection on previously unobserved parent characteristics generates minimal bias in our estimates of the causal impact of scores on later outcomes.

TABLE 6: Correlation Between Teacher Quality Measures

	Baseline	Parent Controls	Lagged Scores	No Controls
	(1)	(2)	(3)	(4)
Baseline	1.0000			
Parent	0.999	1.0000		
Lagged Scores	0.904	0.902	1.0000	
No Controls	0.296	0.292	0.362	1.0000

Notes: This table reports correlations between test score estimates from four models, each using a different control vector. The models are run on a constant sample of 76,879 classrooms observations for which the variables needed to estimate all five models are available. Model 1 uses the baseline control vector defined in the notes to Table 5. Model 2 adds the following parental characteristics to model 1: parent's age at child's birth; mean household income; whether the parent owned a house, invested in a 401k, or was married while child was 19–21; and the parent-child match rate. Model 3 conditions only on lagged scores, using cubics in math and reading interacted with indicators for missing lagged scores. Model 4 includes no controls at all.

IV.B. Impacts of Scores on Outcomes in Adulthood

The results in the previous section show that teacher quality is a plausible instrument for changes in students' test scores. In this section, we analyze the link between scores and long-run outcomes. We do so by regressing outcomes in adulthood Y_i on teacher quality $\hat{\mu}_{j(i,g)}$ and observable characteristics.

We begin by pooling the data across all grade levels and then present results that estimate grade-specific coefficients on teacher quality. Recall that each student appears in our dataset once for every subject-year with the same level of Y_i but different values of $\mu_{j(i,g)}$. Hence, in this pooled regression, the coefficient estimate β represents the mean impact of having higher scores for a single grade between grades 4–8. We account for the repeated student-level observations by clustering standard errors at the school-cohort level as above.

We first report estimates based on comparisons of students assigned to different teachers. We then compare these estimates to those obtained from the "teacher switcher" research design developed by Chetty, Friedman, and Rockoff (2011). This method isolates quasi-experimental variation in teacher quality (and therefore in scores) by looking only at changes in teacher assignment due to changes in the teaching staff within a given school-grade cell. We analyze impacts of scores on three sets of outcomes: college attendance, earnings, and other indicators such as teenage birth rates.[18]

1. College Attendance

We begin by analyzing the impact of scores on college attendance at age 20, the age at which college attendance rates are maximized in our sample. In all figures and tables in this section, we condition on the classroom-level controls used above in equation (2).

Figure 9a plots college attendance rates at age 20 against scores (using teacher assignment as the instrument). Having higher scores in a single grade raises a student's probability of attending college significantly. A 1 SD increase in test scores increases college attendance by 4.92 percentage points at age 20, relative to a mean of 37.8%.

To confirm that the relationship in Figure 9a reflects the causal impact of scores rather than selection bias, we implement tests analogous to those in the previous section. Table 7 presents OLS regression estimates of the impacts of scores. The first column replicates Figure 9a. In Column 2, we replace actual college attendance with predicted attendance based on parent characteristics, constructed in the same way as predicted scores above. The estimates show that one would not have predicted any significant difference in college attendance rates across students with different teacher assignment based on parent characteristics, confirming that selection on observables is minimal.

To address potential bias due to unobservables, we use the cross-cohort teacher switcher identification strategy. We regress changes in mean college attendance rates across adjacent cohorts within a school-grade cell against the change in mean teacher quality $\bar{\mu}_{sgt}$ due to teacher staff changes. Students who happen to be in a cohort in their school in which teachers raise scores are significantly more likely to go to college. The estimate of $\beta = 5.97\%$ from this quasi-experimental variation is very similar to that obtained from the cross-classroom comparison in Column 1, though less precise because it exploits much less variation. The null hypothesis that $\beta = 0$ is rejected with $p < 0.01$, while the hypothesis that β is the same in Columns 1 and 3 is not rejected. We conclude based on this evidence that higher scores generate robust and significant improvements in college attendance rates. Figure 9b shows the impact of scores on college attendance across ages. Although the effect is highest at earlier ages, there is a significant impact through age 25 on the college attendance rate.

Next, we analyze whether test score gains via better teachers also improve the quality of colleges that their students attend. Figure 9c plots this projected earnings-based index of college quality against scores (instrumented using teacher assignment). Again, there is a highly significant relationship between the quality of colleges students attend and the scores they earned in grades 4–8 ($t = 9.5$, $p < 0.0001$). A 1 SD improvement in scores raises college quality by $1,644 (6.6%) on average.

TABLE 7: Impacts of Test Scores on College Attendance

Dependent Variable:	College at Age 20	Predicted College at Age 20	Changes in College Attendance	Changes in Predicted Score	College Quality at Age 20	Changes in College Quality	College Quality High Income	College at Age 25
	(%)	(%)	(%)	(%)	($)	($)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test Scores	4.92	0.463			1,644		3.59	2.75
	(0.65)	(0.261)			(173)		(0.61)	(0.70)
Changes in Mean Test Scores			6.101	0.008		1,319		
			(2.094)	(0.011)		(539)		
Controls	x	x			x		x	x
Source of Variation	X-Class	X-Class	X-Cohort	X-Cohort	X-Class	X-Cohort	X-Class	X-Class
Observations	3,095,822	3,097,322	25,073	25,073	3,095,822	24,296	3,095,822	985,500
Mean of Dependent Variable	37.8	37.8	35.9	0.2	24,815	24,293	19.8	18.1

Notes: Each column reports coefficients from a separate OLS regression. Columns 1, 2, 5, 7, and 8 use cross-class variation, while Columns 3, 4, and 6 use cross-cohort variation. The dependent variable for Column 1 is a dummy variable for college attendance at age 20. The dependent variable for Column 2 is the predicted value generated from a regression of college attendance at age 20 on a quartic in parent mean income (measured during the years when a student is 19–21), a dummy variable for whether the parents are married at some time during the sample, an interaction of the quartic with the married-dummy, a dummy variable for parent home ownership, a dummy variable for the age difference between the claiming parent being less than 20 years, and a dummy variable for parents' contribution to tax-deferred savings accounts. The dependent variable for Column 5 is our earnings-based college-quality measure. The dependent variable in Column 7 is a dummy variable for college-quality being higher than the median college quality among those attending college, which is $29,182. The dependent variable in Column 8 is a dummy variable for college attendance at age 25. All cross-class regressions include the full vector of controls, including a flexible polynomial in lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, teacher experience, and year- and grade-level fixed effects. For the cross-cohort variation in Columns 3, 4, and 6, we use changes in test scores as the main independent variable. For each set of adjacent years, we calculate the raw change in a number of outcome variables that are the dependent variables; each of these differences are cross-cohort changes. We include no controls in the cross-cohort regressions except year fixed effects. We cluster standard errors at the school X-cohort level.

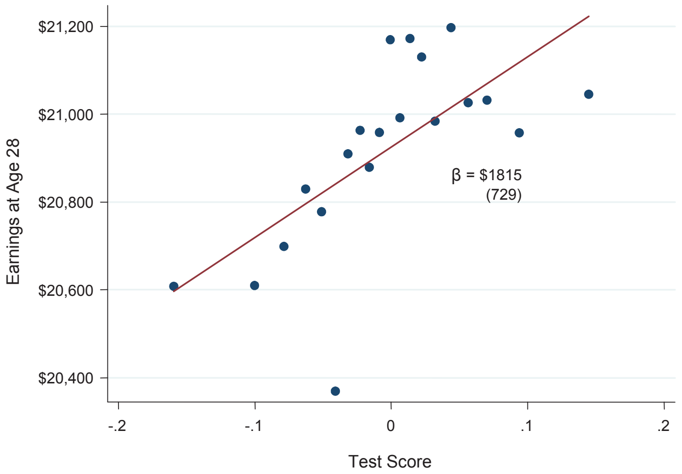
The $1,644 estimate combines intensive and extensive margin responses because it includes the effect of increased college attendance rates on projected earnings. Isolating intensive margin responses is more complicated because of selection bias: students who are induced to go to college because of higher scores tend to attend low-quality colleges, pulling down mean earnings conditional on attendance. To investigate movement on the intensive margin, we define an indicator for "high quality" colleges as those with average earnings above the median among colleges that students attend in our sample, which is $29,182. We regress this indicator on scores (instrumented by teacher assignment as above) in the full sample, including students who do not attend college. Column 7 of Table 7 shows that scores increase the probability that students attend high quality colleges. A 1 SD increase in score raises the probability of attending a high quality college by 3.6%, relative to a mean of 19.8%. This suggests that there is a very large increase in the attendance of high-quality colleges, which is most consistent with an intensive margin effect. Second, we derive a lower bound on the intensive margin effect by assuming that those who are induced to attend college attend a college of average quality. The mean college quality conditional on attending college is $38,623 and the mean quality for those who do not attend college is $16,361. Hence, at most ($38,623 – $16,361) × 4.9% = $1,091 of the $1,644 impact is due to the extensive margin response, confirming that teachers improve the quality of colleges that students attend.

Finally, in Column 8, we analyze college attendance at age 25 instead of 20. Higher scores continue to increase college attendance at age 25, which partly reflects attendance of graduate or professional schools. As expected, the impacts on college attendance at age 25 are smaller in magnitude (2.75% per 1 SD of score) because the mean college attendance rate at age 25 is 18.1% in this sample. We find similar impacts of scores on college attendance at other ages, with magnitudes that decline smoothly with age as expected (not reported). These continued impacts on college attendance in the mid 20's affect our analysis of earnings impacts, which we turn to next.

2. Earnings

The correlation between annual earnings and lifetime income rises rapidly as individuals enter the labor market in their twenties and begins to stabilize only in the late twenties.[19] We therefore begin by analyzing the impacts of scores on earnings at age 28, the oldest age at which we have a sufficiently large sample of students in order to obtain precise estimates. Figure 10 plots earnings at age 28 against scores, instrumenting with teacher assignment and conditioning on the same set of classroom-level controls as above. Earning a higher score has a clear, statistically significant impact on earnings, with the null hypothesis of $\beta = 0$ rejected with $p < 0.01$. A 1 SD increase in scores in a single grade is estimated to increase earnings at age 28 by $1,815, approximately 8.9% relative to mean earnings in this sample of $20,912. This result also appears in Column 1 of Table 8. Column 2 shows the effect on wages at age 30, and the coefficient is slightly larger than that at age 28. But with so few observations the standard error is too large to interpret the coefficient meaningfully, so we restrict to wages up through age 28 for the remainder of the paper.

FIGURE 10: Effect of Test Scores on Earnings at Age 28



Notes: This figure plots the relationship between test scores and wage earnings at age 28. We measure wage earnings from W–2 forms issued by an individual's employer. We calculate residuals for both wage earnings and test scores from a regression on the full control vector, adding back the sample means for ease of interpretation. We group the observations into 20 equal-sized (5 percentile-point) bins based on test scores and plot the average value of both the wage earnings and scores within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. Standard errors are clustered at the school-cohort level. Earnings are in real

To gauge the magnitude of this effect, suppose that the percentage gain in earnings remains constant over the life-cycle and that earnings are discounted at a 3% annual rate back to age 12, the mean age in our sample. Under these assumption, mean NPV earnings in the U.S. population is approximately $522,000.[20] Hence, the financial value of having 1 SD higher scores is $46,190 per grade.

We next analyze how score affects the trajectory of earnings by examining earnings impacts at each age from 20 to 28. We run separate regressions of earnings at each age on score and the standard vector of classroom controls. Figure 11a plots the coefficients from these regressions (which are reported in Appendix Table 9) divided by average wage earnings at each age in our sample vs. age. The impact of scores on earnings rises almost monotonically with age. At early ages, the impact of higher scores is *negative* and significant, which is consistent with our finding that higher scores induce students to go to college. As these students enter the labor force, they have steeper earnings trajectories and eventually earn significantly more than students who had lower scores in grades 4–8. The earnings impacts become positive and statistically significant starting at age 26. By age 28, the earnings impact passes 1% of earnings, as in Figure 11a. Stated differently, scores increase earnings growth rates over the 20s. In Column 2 of Table 8, we regress the change in earnings from age 22 to age 28 on scores. A 1 SD increase in scores is estimated to increase earnings growth by $1,802 over this period.

TABLE 8: Impacts of Test Scores on Earnings

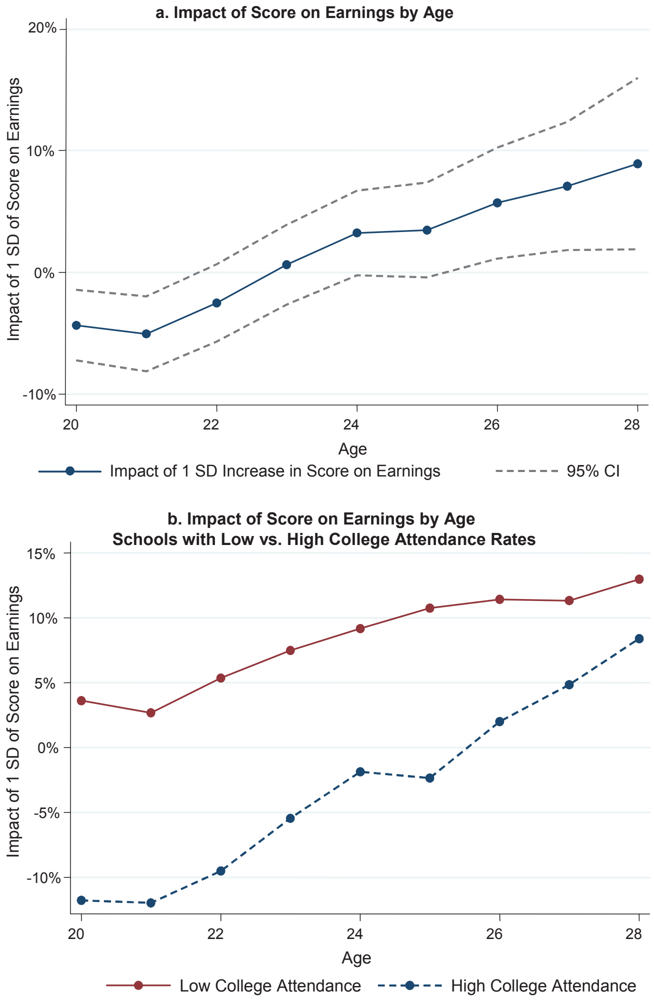
Dependent Variable:	Earnings at Age 28 ($)	Earnings at Age 30 ($)	Wage Growth to Age 28 ($)	College at Age 25 (%)	College at Age 25 (%)	Wage Growth to Age 28 ($)	Wage Growth to Age 28 ($)
School Level College Attendance Rate				Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test Scores	1,815	2,058	1,802	0.526	4.728	1,403	2,838
	(729)	(1,953)	(636)	(0.789)	(1.152)	(661)	(1,118)
Controls	x	x	x	x	x	x	x
Observations	368,427	61,639	368,405	528,065	457,435	201,933	166,472
Mean of Dependent Variable	20,912	22,347	14,039	14.30	22.43	10,159	18,744

Notes: Each column reports coefficients from a separate OLS regression. The dependent variable for Column 1 is wage earnings reported on W–2 forms at age 28. Column 2 repeats column 1 but with wage earnings from W–2 forms at age 30. The dependent variable for Columns 3, 6, and 7 is the change is wage earnings between ages 22 and 28. The dependent variable for Columns 3–4 is a dummy variable for attending college at age 25. All regressions include the full vector of controls, including a flexible polynomial in lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, teacher experience, and year- and grade-level fixed effects. We split the sample in Columns 4 and 5 and then again in Columns 6 and 7 based on the average college attendance rate at each school; those schools with below median attendance rates appears in Column 4, those with above median college attendance rates in Column 5. The median school-average college attendance rate is 35%. We cluster standard errors at the school-cohort level.

We obtain further insight into the role of college in mediating these changes in earnings-trajectories by comparing the impacts of score among students at schools with low and high college attendance rates. We calculate the mean college attendance rate at age 20 in each school and divide the sample into schools above and below the median college attendance rate. In schools with low college attendance rates at age 20, relatively few students are in college at age 25. As a result, changes in score have virtually no impact on college attendance rates at age 25, as shown in Column 3 of Table 8. In contrast, in schools with high college attendance rates in Column 4, a 1 SD increase in scores raises college attendance rates by nearly 5% at age 25. If college attendance masks earnings impacts, one would expect the effects of score on earnings to be smaller at age 25 than at age 28 primarily in high college attendance schools.

This is precisely the pattern observed in the data, as shown in Figure 11b, which plots the effect of score on earnings by age, splitting the sample into school with above- or below-median average college attendance rates. For students unlikely to attend college, a 1 SD increase in test scores has a positive impact on earnings at all ages that rises from 3.6 percentage points at age 20 to nearly 13 percentage points at age 28. In contrast, test scores have a large and significantly negative impact on the wages on students who are likely to attend college up through age 23, which then rises sharply to just 8.4 percentage points by age 28.

FIGURE 11: Effect of Test Scores on Earnings by Age



Notes: Panel a. presents the causal effect of test scores on wages at each age, expressed as a fraction of sample average age-specific wages. Each dot represents the estimated coefficient from a separate regression. To construct Panel a., we regress wages at each age on test scores at the class-level including the full control vector, as in Figure 7. The dashed lines represent the 95% confidence interval, clustered at the school-cohort level. Panel b. reproduces Panel a., splitting the sample into students attending schools with above-median and below-median average college attendance (35%). Each dot represents the coefficient from a regression of wages on test scores, scaled by the average wages for those observations used in the specific regression. For example, the first dot in the upper series is the effect of test scores on wages at age 20 for those student at low-college attendance school, divided by average wages at age 20 for those same students.

To gauge how much further one might expect the earnings impacts to rise over time, we analyze the cross-sectional correlation between test scores and earnings, which we are able to estimate for a longer period because we have test score data starting in 1989, 2 years before we begin to obtain teacher assignment information. Appendix Table 4 lists coefficients from OLS regressions of earnings at each age on test scores. These regressions pool all grades, conditional on the same vector of controls used to estimate equation (2), and use a constant sample of students for whom we observe earnings from ages 20–30 to eliminate cohort effects. The correlation between test scores stabilizes in the early 30's at a level that is roughly 20% higher than at age 28. If the causal impacts of scores track these cross-sectional patterns over time, one would predict a lifetime earnings impact closer to 12.5% rather than 10%. One should be able to obtain a more definitive estimate of the long-term earnings impacts of scores in these data in about 5 years.

Given the available data, we conclude that a 1 SD increase in test scores in a single grade raises earnings in early adulthood by at least 10 percent. Note that the cross-sectional relationship between test scores and earnings reported in Appendix Table 4 implies that a 1 SD increase in test scores is associated with an 11.6%

increase in earnings at age 28. Hence, the causal impact of scores is quite similar to the impact one would have predicted based on the cross-sectional relationship between test scores and earnings. This result aligns with previous evidence that improvements in education raise contemporaneous scores, and then fade out in later scores, only to reemerge in adulthood (Deming, 2009; Heckman et al., 2010c; Chetty et al., 2011). Such findings suggest that education may have long-term impacts through pathways outside academic achievement, such as non-cognitive skills.

As a robustness check, we replicate our analysis using the raw (unshrunk) estimates of teacher quality, in Appendix Table 10. In Columns 1–4, we estimate specifications analogous to the main empirical specification on Section III.C. using OLS. The estimated coefficients are roughly half of those estimated above, reflecting substantial attentuation from measurement error in teacher quality. As an alternative approach to correcting for this measurement error, we then regress each outcome on score using the raw teacher effects $\bar{v}_j$ as instruments; the resulting coefficients appear in Columns 5–7. Here we find very similar results to baseline; none of the three coefficients estimated via instrumental Variables differ by more than half of one standard error from the corresponding baseline result. Appendix Table 7 includes other sets of results to demonstrate the robustness of our baseline estimates to a number of specification choices to ease our computational burden. First, Panels A–C compare our standard errors (clustered at the school-cohort level) to a series of alternatives in the literature. Second, Panel D compares our results to those using alternative control vectors (including individual-level controls). These alternative estimates suggest that the standard errors are slightly too conservative in our baseline specifications.

3. Other Outcomes

We now analyze the impacts of teacher VA on other outcomes, starting with our "teenage birth" measure, which is an indicator for claiming a dependent who was born while the mother was a teenager (see Section III). We first evaluate the cross-sectional correlations between our proxy for teenage birth and test scores as a benchmark. Students with a 1 SD higher test score are 4 percentage points less likely to have a teenage birth relative to a mean of 8% (Appendix Table 3). Conditional on lagged test scores and other controls, a 1 SD increase in test score is associated with a 1.0 percentage point reduction in teenage birth rates. These correlations are significantly larger for populations that have a higher risk of teenage birth, such as minorities and low-income students (Appendix Table 5). These cross-sectional patterns support the use of this measure as an indicator of success in adulthood even though we can only identify teenage births that are claimed as dependents in the tax data.

Column 1 of Table 9 analyzes the impact of test scores on the fraction of female students who have a teenage birth, using teacher assignment as an instrument. Having a 1 SD higher test score in a single year from grades 4 to 8 reduces the probability of a teen birth by 0.99 percentage points, a reduction of roughly 12.5%, as shown in Figure 12a. This impact is again very similar to the cross-sectional correlation between scores and teenage births, echoing our results on earnings and college attendance.

TABLE 9: Impacts of Test Scores on Other Outcomes

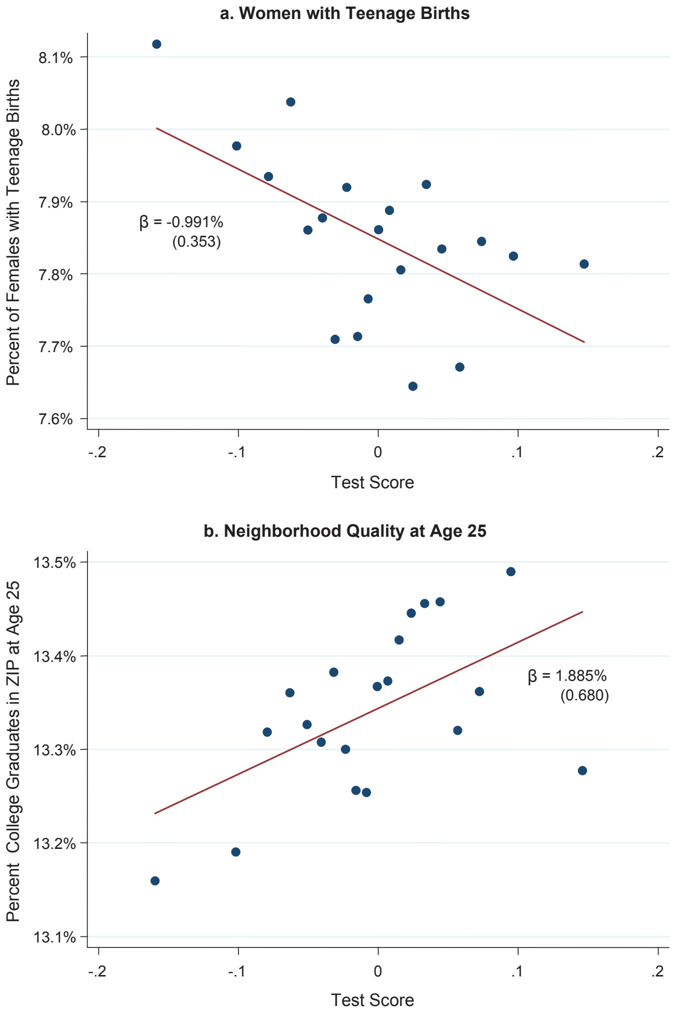
Dependent Variable:	Teenage Birth (%)	Neighborhood Quality at Age 25 (%)	Neighborhood Quality at Age 28 (%)	401(k) at Age 25 (%)	401(k) at Age 25 (%)
School Level College Attendance Rate				Low	High
	(1)	(2)	(3)	(4)	(5)
Test Scores	−0.991	0.628	1.439	1.885	−1.780
	(0.353)	(0.194)	(0.310)	(0.680)	(0.987)
Controls	x	x	x	x	x
Observations	1,826,742	1,168,965	310,638	725,140	646,955
Mean of Dependent Variable	7.9	13.3	13.6	12.1	19.2

Notes: Each column reports coefficients from a separate OLS regression. The dependent variable in Column 1 is a dummy variable for having a teenage birth, which we measure using the claiming of a dependent for tax purposes who is fewer than 20 years younger than the individual. The dependent variable for Columns 2 and 3 is the fraction of residents in an individual's ZIP code of residence with a college degree or higher at ages 25 and 28, respectively, measured from the 2000 Census. We observe ZIP code for either 1040 or W–2 forms filed in the current year, or imputed from past years for non-filers. The dependent variable for Columns 4 and 5 is a dummy variable for whether an individual made contributions to a 401(k) plan at age 25. We split the sample for Columns 4 and 5 by the average college attendance rate of a student's school, as in Table 5. All regressions include the full vector of controls, including a flexible polynomial in lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, teacher experience, and year- and grade-level fixed effects. We cluster standard errors at the school X-cohort level.

Column 2 of Table 9 analyzes the impact of score on the quality of the neighborhood in which students live at age 25, measured by the percent of college graduates living in that neighborhood. A 1 SD increase in score raises neighborhood quality by 0.63 percentage points (5% of the mean) by this metric, as shown in Figure 12b. Column 3 shows that this impact on neighborhood quality more than doubles at age 28, consistent with the growing earnings impacts documented above.

Finally, we analyze impacts on retirement savings. Score does not have a significant impact on 401(k) savings at age 25 in the pooled sample (not reported). However, Column 4 shows that for students in schools with low levels of college attendance, a 1 SD increase in score raises the probability of having a 401(k) at age 25 by 1.9 percentage points (16% of the mean). In contrast, Column 5 shows that for students in high

FIGURE 12: Effects of Test Scores on Other Adult Outcomes



Notes: Panel a. plots the relationship between teenage births and test scores. Panel b. plots the relationship between neighborhood quality at age 25 and test scores. We define a teenage birth as an individual claiming a child fewer than 20 years younger as a dependent on the 1040 tax form in any year in our sample. We define neighborhood quality as the fraction of residents within ZIP code with a college degree for the ZIP code from which the individual filed the 1040. For individuals that do not file, we use the ZIP code of residence reported on a W–2 form. For individuals with no W–2, we impute forward from the last year with ZIP code present. For both panels, we calculate residuals for both the dependent variable of interest and test scores from a regression on the full control vector, adding back the sample means for ease of interpretation. We group the x-axis residuals into 20 equal-sized (5 percentile-point) bins based on the test scores and plot the mean value of the y-variable residuals vs. the mean value of x-variable residuals within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. Standard errors are clustered at the school-cohort level.

college-attendance schools, the point estimate of the impact is negative. These results are consistent with the impacts on earnings trajectories documented above. In schools with low college attendance rates, students with high scores find better jobs by age 25 and are more likely to start saving in 401(k)s. In schools with high college attendance rates, students with high scores are more likely to be in college at age 25 and thus may not begin saving for retirement till a later age.

4. Heterogeneity Analysis

In Table 10, we analyze whether improvements in scores have heterogeneous effects across demographic groups and subjects. We study impacts on college quality at age 20 because the heterogeneity analysis requires large samples and because the college quality measure provides a quantitative metric based on projected earnings gains. Each number in the table is from a separate regression of college quality on scores, with the same classroom-level controls as in the previous sections. Columns 1 and 2 consider heterogeneity by gender, while Columns 3 and 4 consider heterogeneity by parent income, dividing students into groups above and below the median level of parent income in the sample. The first row of the table pools both subjects, while the second and third rows consider math and English scores separately.[21]

TABLE 10: Heterogeneity in Impacts of Test Scores

	Panel A: Impacts of Test Score by Demographic Groups					
	Girls	Boys	Low Income	High Income	Minority	Non-Minority
	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: College Quality at Age 20					
Math and Reading Scores	1,903	1,386	1,227	2,087	1,302	2,421
	(211)	(203)	(174)	(245)	(154)	(375)
	Panel B: Impacts of Test Score on College Quality: Math vs. Reading					
	Elementary School			Middle School		
	(1)	(2)	(3)	(4)	(5)	(6)
Math Score	1,095		638	1,648		1,374
	(176)		(219)	(357)		(347)
Reading Score		1,901	1,281		2,896	2,543
		(303)	(376)		(586)	(574)
Joint Controls	x	x	x	x	x	x

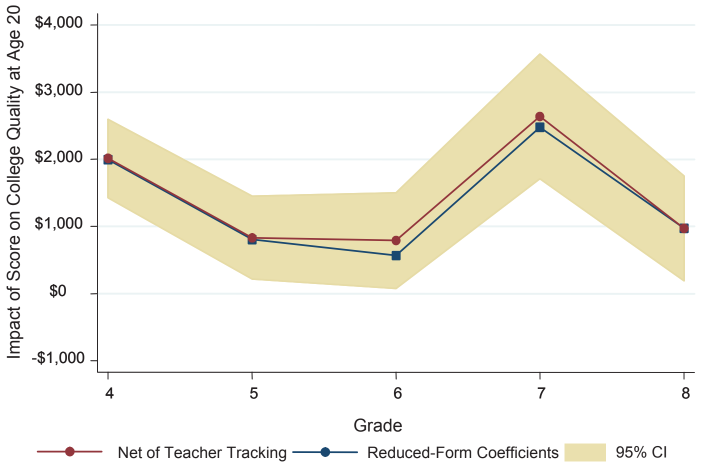
Notes: Each cell reports a coefficient from a separate OLS regression of an outcome on test scores, using teacher assignment as an instrument. The dependent variable is the earnings-based college-quality measure (see Table 1 for a detailed description). All regressions include a partial vector of controls, including a flexible polynomial in lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, year- and grade-level fixed effects, and teacher experience dummies. In Panel A, we split the sample in Columns 1 and 2 between boys and girls. We split the sample in Columns 3 and 4 based on whether a student's parental income is higher or lower than median in sample, which is $39,835. We split the sample in Columns 5 and 6 based on whether a student belongs to an ethnic minority. In Panel B, we split the sample into elementary schools (schools where the student is taught by the same teacher for both math and English) and middle schools (which have different teachers for each subject). We then regress college quality on math teacher quality, English teacher quality, and both teacher qualities jointly, controlling for the full set of class-level controls in each classroom. We cluster standard errors at the school-cohort level.

Two lessons emerge from Panel A of Table 10. First, the point estimates of the impacts of teacher VA are slightly larger for girls than boys, although one cannot reject equality of the impacts at conventional significance levels. Second, the impacts are smaller for lower-income and minority households in absolute terms. For instance, a 1 SD increase in scores raises college quality by $1,227 for children whose parents have below-median income, compared with $2,087 for those whose parents have above-median income. However, the impacts are roughly constant as a percentage of mean college quality: 5.6% for low-income students vs. 7.5% for high-income students.

Panel B of Table 10 analyzes differences in teachers' impacts across subjects. For these regressions, we split the sample into elementary school and middle school classrooms. The first three columns consider the effects in elementary school, in which students have one teacher for all subjects. The second set of three columns considers middle school classrooms, when students have different teachers for each subject. We first show the effects of each measure separately (but on a constant sample), and then combine the two measures into a single regression.

Comparing across subjects, the coefficients on test score are generally larger in English than math. Including both reading and math teacher quality has very different effects in elementary vs. middle school. In elementary school, students have one teacher for all subjects; math and reading quality are therefore highly correlated ($r = 0.59$), since they are just different measures of quality for the same teacher. Not surprisingly, the magnitude of the coefficients drops by nearly 40% when included together in the regression in Column 3. In contrast, students usually have different teachers for math and reading in middle school, and the correlation between the quality measures is much lower ($r = 0.17$). Thus, including the two measures together in Column 6 reduces the point estimates by only 10%.

FIGURE 13: Impacts of Test Scores on College Quality by Grade



Notes: This figure plots the impact of a 1 SD increase in test scores in each grade from 4–8 on our earnings-based index of college quality (projected earnings at age 30 based on the college in which the student is enrolled at age 20). Each point in the upper series represents the coefficient on test scores from a separate regression of college quality at age 20 on test scores including the standard control vector. The shaded area represents a 95% confidence interval clustered on school-cohort. These coefficients represent the reduced-form effect of improved teacher quality in each grade, including not only the direct impact on earnings but also the indirect effect through improved teacher quality in future years.

Figure 13 displays the impact of test scores on long-run outcomes across grades, and the coefficients are reported in Appendix Table 1. We use college quality (projected earnings at age 30 based on college enrollment at age 20) as the outcome in order to have adequate precision to identify grade-specific effects. We estimate the coefficients using specifications analogous to Column 5 of Table 7 for each grade separately, using the subsample of students who are in our school district dataset for all grades from 4–8. The long-term impacts of score are relatively stable across grades. Although the estimates in each grade have relatively wide confidence intervals, there is no systematic trend in the impacts. No one grade has a significantly larger impact than any other grade. This pattern is consistent with the cross-sectional correlations between test scores and adult outcomes, which are also relatively stable across grades (Appendix Table 6).

V. Conclusion: Long-Term Effects of Tax Credits

We now combine the evidence from our preceding analysis to answer our original question: what are the long-term impacts of tax credits on earnings? Our estimates from the first part of our analysis imply that a $1,000 tax credit increases a child's test score by 6% of a standard deviation (taking the more conservative estimate for reading scores from Table 2). This is a relatively large effect; for comparison, the standard deviation of teacher impacts on achievement is approximately 10% of a standard deviation, which is similar to the estimates of Dahl and Lochner (2011).

These score gains themselves have no direct economic interpretation, as we do not know how test score gains translate into earnings gains. Ideally, we would directly analyze the long-term impacts of the EITC or CTC on children's future earnings, but our data do not cover a long enough time period to permit such an analysis. As a feasible alternative, we evaluated the effects of a different intervention—better teachers—to understand how test score gains translate into earnings gains. Under the assumption that score increases generated from these different policies have the same long-run effects, the research in this paper allows us to achieve the original objective of understanding the long-run impacts of cash grants through tax policy on children's long-run outcomes.

Our estimates imply that a 1 SD increase in test scores raises college attendance rates by 5.1 percentage points. Putting this estimate together with our estimate of the impact of tax credits on test scores, we estimate that a $1,000 tax credit to families with young children would increase the probability of college attendance rates at age 20 by 4.9 × 0.06 = 0.294 percentage points relative to a base of 36.2%.

While the impact on college attendance may appear modest relative to other tax policies targeted at raising college attendance rates, a tax credit to families with young children generates a significant dollar earnings gain over a student's lifetime. We estimate that a 1 SD increase in test scores raises earnings by approximately 9 percentage points. Hence, a $1,000 tax credit would raise a child's lifetime earnings by 0.09 × 0.06 = 0.54 percentage points. The dollar gains in lifetime earnings are of the same order of magnitude as the cost of the tax credit because a small percentage increase in earnings over a lifetime adds up to a large sum in present value.

These findings suggest that there are substantial returns to public policies that help poor families with children. Consider, for instance, the expansion of the EITC in 2009 to pay an additional credit to families with 3 children. This policy was passed in 2009 as part of the American Recovery and Reinvestment Act for 2 years, and the 2011 budget made this change permanent. Specifically, this change raised the maximum credit for families with 3 or more children by 5% from $5,028 to $5,657. These results suggest that this reform may have increased the NPV earnings of children of these families by more than 5%.

Although this analysis has used the federal EITC for identification, the findings apply equally to state EITC programs. Many states offer an EITC that is defined as a percentage of the federal credit; in 2009, this percentage varies from 0% in 28 states to as much as 43% (for families with 3 children in Wisconsin). But these programs have come under pressure as states grapple with declining revenue streams. This constraint is especially difficult in states with a balanced budget requirement. For instance, Michigan planned to increase the state EITC from 10% to 20% in 2009, but the legislature froze the credit at 10% of the federal EITC. This freeze saved the state about $100 million but also deprived children from poor families of more than $100 million in NPV earnings. These gains, which come far in the future, are often difficult to represent vividly in public debates.

Many states also wrestle with the choice to make state EITC credits refundable. Of the earned income credits in the 22 states plus the District of Columbia, all but four are fully refundable. (The credits in Delaware and Virginia are not refundable, while those in Maine and Rhode Island are partially refundable). Credits that are not refundable hit those families with the lowest income, and therefore the least tax to offset. The impacts on long-term outcomes are likely highest among these poorest families because of credit constraints. Economic theory suggests that families will invest in the education of their children, for instance by moving to better school districts, when the returns to these investments are large. However, families cannot easily borrow to pay for primary or secondary education. States should therefore carefully consider the benefits and costs of refundability before restricting access to the EITC for families with few tax liabilities to offset.

There are many caveats that one must keep in mind when interpreting these results. First, our estimates of the impacts of tax credits on test scores rely on very strong assumptions on the cross-sectional pattern of test scores and household income. This relationship must hold constant across high and low-income households in order for our identification strategy to be valid. There are a number of reasons that this might not be the case. For instance, patterns of mean-reversion in income may imply that the permanent income of low-income families is substantially higher relative to current income than for higher income families. Children of low-income families may also face more heterogeneous school or neighborhood conditions that could attenuate an otherwise stable relationship. Future research is necessary to determine whether these strong assumptions are met.

Second, the long-run impact of test score increases from different sources may vary considerably. In the extreme, increased student test scores through cheating should not have any long-run impact (and might even have a negative one). In this case, we must assume that the long-run effect of a higher score that comes from a better teacher is the same as that from an increase in tax credits for a child's household. In order to directly evaluate this assumption, one must examine the long-run outcomes of the same students whose households received the large tax credits. Unfortunately our data do not span enough years to make this exercise feasible. But examining the direct impacts of tax credits on long-run outcomes should be a first-order priority for research when these data become available.

Finally, our results do not shed light on the mechanism through which an increase in tax credits aids student achievement. Are families directly spending more on school-related resources for children? Are families moving into neighborhoods with better schools and better peers? In order to have confidence that our estimates truly reflect the impact of tax credits, future research should investigate this mechanism. Because of these limitations, policy makers should exercise caution when extrapolating evidence from the tax credits we have studied to predict the likely impact of future credits. The most important lesson of our analysis is that tax policy could have substantial long-term impacts and that future research should focus on analyzing these issues further.

Endnotes

[1] The tax data were accessed through contract TIRNO-09-R-00007 with the Statistics of Income (SOI) Division at the U.S. Internal Revenue Service. Sarah Griffis, Jessica Laird, and Heather Sarsons provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard and the National Science Foundation is gratefully acknowledged.

[2] All tests were administered in late April or May during the early-mid 1990s, and students were typically tested in all grades on the same day throughout the district. Statewide testing dates varied to a greater extent, and were sometimes given earlier in the school year (e.g., February) during the latter years of our data.

[3] The standard deviation of 4th and 8th grade reading and math achievement in this district ranges from roughly 95% to 105% of the national standard deviation on the National Assessment of Educational Progress, based on data from 2003 and 2009, the earliest and most recent years for which NAEP data are available. Mean scores in the district are significantly lower than the national average, as expected given the urban setting of the district.

[4] 5% of students switch classrooms or schools in the middle of a school year. We assign these students to the classrooms in which they took the test in order to obtain an analysis dataset with one observation per student-year-subject. However, when defining class and school-level means of student characteristics (such as fraction eligible for free lunch), we account for such switching by weighting students by the fraction of the year they spent in that class or school.

[5] We obtain similar results using household adjusted gross income reported on individual tax returns. We focus on the W–2 measure because it provides a consistent definition of individual wage earnings for both filers and non-filers. One limitation of the W–2 measure is that it does not include self-employment income.

[6] Colleges are not required to file 1098–T forms for students whose qualified tuition and related expenses are waived or paid entirely with scholarships or grants; however, the forms are generally available even for such cases, perhaps because of automated reporting to the IRS by universities.

[7] See Chetty et al. (2011) for a comparison of total enrollment based on 1098–T forms and statistics from the Current Population Survey. Chetty et al. use this measure to analyze the impacts of Project STAR on college attendance. Dynarski et al. (2011) show that using data on college attendance from the National Clearinghouse yields very similar estimates to Chetty et al.'s findings, providing further confirmation that the 1098–T based college indicator is accurate.

[8] For the small fraction of students who attend more than one college in a single year, we define college quality based on the college that received the largest tuition payments on behalf of the student.

[9] An alternative matching procedure would be to omit students in years when they are not claimed by any household. We have conducted robustness checks using this procedure and find nearly identical results.

[10] Note that AGI is not technically sufficient for determining credit eligibility. For instance, taxpayers may choose to include non-taxable combat pay in the income on which their EITC payment is calculated. Nevertheless, AGI is a sufficiently close approximation for most households.

[11] For instance, suppose Household A files in year t with two dependents aged 7 and 10. If that household does not file in year $t + 1$, we impute that the household included two dependents aged 8 and 11.

[12] We define the mother's age at child's birth as missing for 471 observations in which the implied mother's age at birth, based on the claiming parent's date of birth, is below 13 or above 65. These are typically cases where the parent does not have an accurate birth date recorded in the SSA file.

[13] This college enrollment rate is slightly higher than for the U.S. population, perhaps reflecting the wide availability of community colleges in this urban school district.

[14] The EITC schedule differs by filing status. Figure 1a shows the schedule for single filers.

[15] All figures are quoted in 2010 dollars.

[16] In tax years after 2008, the EITC included additional payments to households claiming three or more children; since our sample period ends in 2008, this recent reform is not relevant for our analysis.

[17] The standard error is slightly lower than before, though it is the same when rounded for significant digits.

[18] One observable characteristic of teachers that predicts VA is teacher experience. Using a fixed-effects specification analogous to Rockoff (2004), we find in our data that students assigned to rookie teachers have 0.03 SD lower test score gains. Because the impact on scores is small, we have insufficient power to test whether these inexperienced teachers have impacts on earnings and other adult outcomes that are commensurate to the impacts of teacher VA.

[19] Although individuals' earnings trajectories remain quite steep at age 27, earnings levels at age 27 are highly correlated with earnings at later ages (Haider and Solon, 2006), a finding we confirm in the tax data (Chetty et al., 2011, Appendix Table I).

[20] We use the mean wage earnings of a random sample of the U.S. population in 2007 to obtain a baseline earnings profile over the lifecycle (see Chetty et al., 2011, for details).

[21] Note that in elementary school, the same teacher typically teaches both math and reading, and VA in math is correlated with VA in reading.

References

Aaronson, Daniel; Lisa Barrow; and William Sander (2007). "Teachers and Student Achievement in Chicago Public High Schools." *Journal of Labor Economics* 24:1, 95–135.

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

Altonji, Joseph; Todd Elder; and Christopher Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–184.

American Community Survey, (http://www.census.gov, U.S. Census Bureau), 2006–2008 ACS 3-year data.

Baker, Eva L., Paul E. Barton; Linda Darling-Hammond; Edward Haertel; Helen F. Ladd; Robert L. Linn; Diane Ravitch; Richard Rothstein; Richard J. Shavelson; and Lorrie A. Shepard (2010). "Problems with the Use of Student Test Scores To Evaluate Teachers." *Economic Policy Institute Briefing Paper #278*, August.

Barlevy, Gadi; and Derek Neal (Forthcoming). "Pay for Percentile." *American Economic Review*.

Boyd, Donald; Pamela Grossman; Hamilton Lankford; Susanna Loeb; and James Wyckoff (2009). "Who Leaves? Teacher Attrition and Student Achievement." *CALDER Center Working Paper #23*, March.

Cameron, Colin A.; Jonah B. Gelbach; and Douglas Miller (2011). "Robust Inference with Multi-way Clustering." *Journal of Business and Economic Statistics,* 29 (2): 238-249.

Carrell, Scott E., and James E. West (2010). "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy*, Vol. 118, No. 3 (June), pp. 409–432.

Chetty, Raj; John N. Friedman; Nathaniel Hilger; Emmanuel Saez; Diane Whitmore Schanzenbach; and Danny Yagan (Forthcoming). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics*.

Chetty, Raj, and John N. Friedman (2012). "New Evidence on the Long-Term Impacts of Tax Credits." *National Tax Association Proceedings*.

Chetty, Raj; John N. Friedman; and Jonah E. Rockoff (2011). "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." *Harvard University Working Paper*.

Corcoran, Sean P. (2010). "Can Teachers be Evaluated by Their Students' Test Scores? Should they Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice." Report for the Annenberg Institute for School Reform, Education Policy for Action Series, September.

Cunha, Flavio; James J. Heckman (2010). "Investing in Our Young People." *NBER Working Paper 16201*, July.

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

Currie, Janet (2010). "Inequality at Birth: Some Causes and Consequences." *Columbia University Working Paper*.

Dahl, Gordon B., and Lance Lochner (2011). "The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit." *American Economic Review*.

Deming, David (2009). "Early Childhood Intervention and Life-Cycle Development: Evidence from Head Start." *American Economic Journal: Applied Economics*, Vol 1, Issue 3, pp. 111–134.

Duflo, Esther (2000). "Child Health and Household Resources in South Africa: Evidence from the Old Age Pension Program." *The American Economic Review*, 90(2): 393–398.

Dynarski, Susan; Joshua M. Hyman; Diane Whitmore Schanzenbach (2011). "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion." *NBER Working Paper 17533*, October.

Gertler, Paul (2004). "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." 94 (2): 336–341.

Goldhaber, Dan and Michael Hansen (2010). "Using Performance on the Job To Inform Teacher Tenure Decisions." *American Economic Review*, 100(2): 250–255.

Gordon, Robert; Thomas J. Kane; and Douglas O. Staiger (2006). "Identifying Effective Teachers Using Performance on the Job." Hamilton Project Discussion Paper 2006-01, April.

Haider, Steven, and Gary Solon (2006). "Life-cycle Variation in the Association Between Current and Lifetime Earnings." *The American Economic Review,* 96, 1308–1320.

Hanushek, Eric A. (1971). "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review*, May (Papers and Proceedings), 61(2), pp. 280–88.

Hanushek, Eric A. (2009). "Teacher Deselection." in *Creating a New Teaching Profession*, ed. Dan Goldhaber and Jane Hannaway, 165–80. Washington, DC: Urban Institute Press.

Hanushek Eric A.; John F. Kain; and Steven G. Rivkin (2004). "Why Public Schools Lose Teachers." *Journal of Human Resources* Vol. 39, No. 2 (Spring), pp. 326–354.

Heckman, James (2000). "Policies to Foster Human Capital." *Research in Economics*, 54(1): 3–56.

Heckman, James J.; Seong H. Moon; Rodrigo Pinto; Peter A. Savelyev; and Adam. Yavitz (2010a). "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics*.

Heckman, James J.; Seong H. Moon; Rodrigo Pinto; Peter A. Savelyev; and Adam Yavitz (2010b). "The Rate of the Return to the High/Scope Perry Preschool Program." *Journal of Public Economics,* 94, 114–128.

Heckman, James; Lena Malofeeva; Rodrigo Pinto; and Peter Savelyev (2010c). "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." unpublished manuscript, University of Chicago.

Internal Revenue Service (2010). *Document* 6961*: Calendar Year Projections of Information and Withholding Documents for the United States and IRS Campuses* 2010–2018, IRS Office of Research, Analysis, and Statistics, Washington, D.C.

Jacob, Brian A (2005). "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89:5–6, 761–796.

Jacob, Brian A., and Steven D. Levitt (2003). "Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating." *The Quarterly Journal of Economics* 118(3): 843-877.

Jacob, Brian A., and Jens Ludwig (2007). "The Effects of Housing Vouchers on Children's Outcomes." University of Chicago Working Paper.

Kane, Thomas, and Douglas Staiger (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *NBER Working Paper No. 14607*.

Jacob, Brian A.; Lars Lefgren; and David P. Sims (2010). "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources*, 45:4, 915–943.

Jacob, Brian A., and Jonah E. Rockoff (2011). "Organizing Schools To Improve Student Achievement: Start Times, Grade Configurations, And Teaching Assignments" Hamilton Project Discussion Paper 2011-08, September.

Jackson, C. Kirabo (2010). "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence From Teachers." *NBER Working Paper No. 15990*, May.

Jackson, C. Kirabo, and Elias Bruegmann (2009). "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics* 1(4): 85–108.

Kane, Thomas J., and Douglas O. Staiger (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *NBER Working Paper No. 14607*.

Kane, Thomas J.; Jonah E. Rockoff; and Douglas O. Staiger (2008). "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," *Economics of Education Review* 27: 615–631.

Kane, Thomas J.; Eric S. Taylor; John H. Tyler; and Amy L. Wooten (2011). “Identifying Effective Classroom Practices Using Student Achievement Data.” *Journal of Human Resources* Summer, Vol. 46, No. 3, 587–613.

Krueger, Alan B (1999). “Experimental Estimates of Education Production Functions.” *Quarterly Journal of Economics*, May, Vol. 114, No. 2, Pages 497–532.

Lockwood, J.R., and Daniel F. McCaffrey (2009). “Exploring Student-Teacher Interactions in Longitudinal Achievement Data.” *Education Finance and Policy* 4(4): 439–467.

McCaffrey, Daniel F.; Tim R. Sass; J.R. Lockwood; Kata Mihaly (2009). “The Intertemporal Variability of Teacher Effect Estimates.” *Education Finance and Policy* 4(4): 572–606.

Morris, Carl (1983). “Parametric Empirical Bayes Inference: Theory and Applications.” *Journal of the American Statistical Association*, 78: 47–55.

Murnane, Richard J. (1975). *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Ballinger.

Neal, Derek A., and Diane Whitmore Schanzenbach (2010). “Left Behind by Design: Proficiency Counts and Test-Based Accountability.” *Review of Economics and Statistics*, May, Vol. 92, No. 2, Pages 263–283.

Oreopoulos, Philip (2006). “Estimating Average and Local Average Treatment Effects of Education When Compulsory School Laws Really Matter.” *American Economic Review* 96:1, 152–175.

Oreopoulos, Philip, and Kjell G. Salvanes (2010). “Priceless: The Nonpecuniary Benefits of Schooling.” *Journal of Economic Perspectives*, 25(1): 159–84.

Rivkin, Steven. G.; Eric. A. Hanushek; and John F. Kain (2005). “Teachers, Schools and Academic Achievement.” *Econometrica,* 73, 417–458.

Rockoff, Jonah (2004). “The Impact of Individual Teachers on Student Achievement.” *The American Economic Review, Papers and Proceedings*, 94(2), pp. 247–252.

Rockoff, Jonah E.; Douglas O. Staiger; Thomas J. Kane; and Eric S. Taylor (Forthcoming). “Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools.” *American Economic Review*.

Rockoff, Jonah E., and Cecilia Speroni (2011). “Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City.” *Labour Economics,* 18, 687–696.

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

Rubin, Donald B.; Elizabeth A. Stuart; and Elaine L. Zanutto (2004). “A Potential Outcomes View of Value-Added Assessment in Education.” *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, Value-Added Assessment Special Issue (Spring), pp. 103–116.

Saez, Emmanuel; Joel B. Slemrod; and Seth H. Giertz (2012). “The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review.” *Journal of Economic Literature*.

Staiger, Douglas O., and Jonah E. Rockoff (2010). “Searching for Effective Teachers with Imperfect Information.” *Journal of Economic Perspectives,* 24, 97–117.

Todd, Petra E., and Kenneth I. Wolpin (2003). “On the Specification and Estimation of the Production Function for Cognitive Achievement.” *The Economic Journal,* Vol. 113, No. 485, Features (February), pp. F3–F33.

U.S. Census Bureau (2010). “School Enrollment—Social and Economic Characteristics of Students: October 2008, Detailed,” Washington, D.C., (http://www.census.gov/population/www/socdemo/school.html).

Appendix A: Matching Procedures

Step 1 [Date of Birth, Gender, Lastname]: We begin by matching each individual from the school-district data to Social Security Administration (SSA) records. We match individuals based on exact date of birth, gender, and the first four characters of last name. We only attempt to match individuals for which the school records include a valid date of birth, gender, and at least one valid last name. SSA records all last names ever associated in their records with a given individual; in addition, there are as many as three last names for each individual from the school files. We keep a potential match if any of these three last names match any of the last names present in the SSA file.

Step 2 [Rule Out on First Name]: We next check the first name (or names) of individuals from the school records against information from W–2 and other information forms present in the tax records. Since these files reflect economic activity usually after the completion of school, we use this information in Step 2 only to "rule out" possible matches. That is, we disqualify potential matches if there exists no first name match between any of the first names on the information returns and any of the first names present in the school data. As before, we conduct these matches on the basis of the first four characters of a first name. For many potential matches we find no first name information in the records; at this step we retain these potential matches. After removing potential matches that are actively mismatched on first name, we then look for students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality (MQ) of matches identified at this stage as $MQ = 1$.

Step 3 [Dependent ZIP code]: For each potential match that remains, we find the household that claimed the individual as a dependent (if the individual was claimed at all) in each year. We then match the location of the claiming household, identified by 5-digit ZIP code, to the home address ZIP code recorded in the school files. We classify potential matches based on the best ZIP code match across all years using the following tiers: exact match, match within 10 (i.e., 02139 and 02146 qualify), match within 100, and non-match. We retain potential matches only that sit in the best available tier of ZIP code match quality. For example, suppose there are 5 potential matches for a given individual, and that there are no exact matches, two matches within 10, two matches within 100, and one non-match. We would retain only the two that matched within 10. We have tried an alternative version of this tier system in which we replace the "within 100" tier with being within the MSA, and the resulting matches are nearly identical. We then look for students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality (MQ) of matches identified at this stage as $MQ = 2$.

Step 4 [Place of Birth]: For each potential match that remains, we match the state of birth from the school records with the state of birth as identified in SSA records. We classify potential matches into three groups: state of birth matches, state of birth does not match but the SSA state is the state of school attendance, and mismatches. Note that we include the second category primarily to account for the immigrants in the school data for whom the recorded place of birth is outside the country. For such children, the SSA state-of-birth corresponds to the state in which they received the social security number, which is often the first state in which they lived after coming to the country. We retain potential matches only that sit in the best available tier of place-of-birth match quality. We then look for students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality (MQ) of matches identified at this stage as $MQ = 3$.

Step 5 [Rule In on First Name]: After exhausting other available information, we return to the first name. To recall, in step 2 we retained potential matches that either actively matched on first name or for which there was no first name available. In this step, we retain only potential matches that actively match on first name, if such a potential match exists for a given student. In addition, we include some information on first name present on 1040 forms filed by potential matches as adults. We then look for students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality (MQ) of matches identified at this stage as $MQ = 4$.

Step 6 [Fuzzy Date-of Birth]: In previous work, we found that 2–3% of individuals had a reported date of birth that was incorrect. In some cases the date was off only by a few days; in others the month or year was off by one, or the transcriber transposed the month and day. To account for this possibility, we take all individuals for whom no eligible matches remained after step 2. (Note that if any potential matches remained after step 2, then we would either settle on a unique best match in the steps that follow or find multiple potential matches even after step 5.) We then repeat step 1, matching on gender, first four letters of last name, and fuzzy date-of-birth. We define a fuzzy DOB match as one where the absolute value of the difference between the DOB reported in the SSA and school data was in the set {1,2,3,4,59,10,18,27} in days, the set {1,2} in months, or the set {1} in years. We then repeat steps 2 through 5 exactly as above to find additional matches. We classify matches found using this fuzzy-DOB algorithm as $MQ = 5.X$, where X is the corresponding MQ from the non-fuzzy DOB algorithm. For instance, if we find a unique fuzzy-DOB match in step 3 using dependent ZIP codes, then $MQ = 5.2$.

The following table shows the distribution of match qualities. In all, we match 85.36% of the students and 88.6% of the student-school year observations.

Match Quality	Frequency	Percent	Cumulative Match Rate (percent)
1	1,098,288	47.28	47.28
2	737,080	31.73	79.01
3	83,315	3.59	82.60
4	28,723	1.35	83.95
5.1	30,367	1.31	85.25
5.2	1,966	0.08	85.34
5.3	371	0.02	85.35
5.4	200	0.01	85.36
Multiple Matches	160,351	6.90	
No Matches	179,687	7.74	

APPENDIX TABLE 1: Impacts of Test Score on College Quality by Grade

	College Quality at Age 20				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	(1)	(2)	(3)	(4)	(5)
Test Score	2,011	832	788	2,638	970
	(296)	(314)	(363)	(472)	(398)

Note: This table replicates the results in Column 5 of Table 7, cutting by grade.

APPENDIX TABLE 2: Impacts of Test Scores on Earnings by Age

	Earnings ($)								
Dependent Variable:	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Test Scores	−211	−322	−211	71	449	558	1,021	1,370	1,815
	(72)	(100)	(136)	(190)	(247)	(319)	(416)	(517)	(729)
Controls	x	x	x	x	x	x	x	x	x
Mean Earnings	4,872	6,378	8,398	11,402	13,919	16,071	17,914	19,322	20,353

Notes: Each column replicates the specification in Column 1 of Table 8, replacing the dependent variable with wages at a given age. In each regression we include observations from all grades. The mean earnings represents the average of the dependent variable for the observations in the estimation sample. All standard errors are clustered by school-cohort.

APPENDIX TABLE 3: Cross-Sectional Correlations Between Outcomes in Adulthood and Test Scores

Dependent Variable:	Earnings at Age 28 ($)	College Quality at Age 20 ($)	College at Age 20 (%)	Teenage Birth (%)	Percent College Grads in Neighborhood at Age 25 (%)
	(1)	(2)	(3)	(4)	(5)
	Panel A: Full Controls				
Test Score	2,539	2,009	5.66	−1.03	0.37
	(76)	(13)	(0.05)	(0.04)	(0.01)
	Panel B: No Controls				
Test Score	7,601	6,030	18.33	−3.84	1.85
	(28)	(6)	(0.02)	(0.02)	(0.01)
	Panel C: Math, Full Controls				
Test Score	2,813	2,131	5.97	−0.88	0.34
	(104)	(18)	(0.07)	(0.06)	(0.02)
	Panel D: English, Full Controls				
Test Score	2,194	1,843	5.27	−1.21	0.38
	(112)	(18)	(0.07)	(0.06)	(0.02)
Mean of Dependent Variable	20,867	24,678	37.17	8.25	13.18

Notes: Each column reports coefficients from a separate OLS regression. The dependent variable for Column 1 is wage earnings reported on W–2 forms at age 28. The dependent variable for Column 2 is our earnings-based college-quality measure. The dependent variable for Column 3 is a dummy variable for college attendance at age 20. The dependent variable in Column 4 is a dummy variable for having a teenage birth, which we measure using the claiming of a dependent for tax purposes who is fewer than 20 years younger than the individual. The dependent variable for Column 5 is the fraction of residents in an individual's ZIP code of residence with a college degree or higher at age 25, measured from the 2000 Census. We observe ZIP code for either 1040 or W–2 forms filed in the current year, or imputed from past years for non-filers. Regressions include the full vector of controls, including a flexible polynomial in lagged student score and the average lagged score of other students in the class, student characteristics, class-level characteristics, school-grade level average lagged test scores and characteristics, teacher experience, and year- and grade-level fixed effects.

APPENDIX TABLE 4: Correlation Between Test Scores and Earnings by Age

	Earnings ($)										
Dependent Variable:	20	21	22	23	24	25	26	27	28	29	30
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
No Controls	435	548	1,147	2,588	3,516	4,448	5,507	6,547	7,440	8,220	8,658
	(15)	(19)	(24)	(30)	(36)	(42)	(49)	(56)	(63)	(68)	(72)
Controls	178	168	354	942	1,282	1,499	1,753	2,151	2,545	2,901	3,092
	(46)	(60)	(75)	(94)	(111)	(130)	(151)	(172)	(191)	(208)	(219)
Mean Earnings	4,093	5,443	6,986	9,216	11,413	13,811	16,456	19,316	21,961	23,477	23,856
Effect as a % of Mean Earnings	4.4%	3.1%	5.1%	10.2%	11.2%	10.9%	10.7%	11.1%	11.6%	12.4%	13.0%

Notes: These coefficients represent the estimates on scores in a cross-sectional regression of wages at age 28 on score. In each regression we include observations from all grades. The mean earnings represents the average of the dependent variable for the observations in the estimation sample. In each regression we include observations from all grades. The first set of regressions does not include controls. The second set of regressions includes the full control vector. All standard errors are clustered by school-cohort.

APPENDIX TABLE 5: Heterogeneity in Cross-Sectional Correlations Across Demographic Groups

Dependent Variable:	Earnings at Age 28 ($)	College at Age 20 (%)	College Quality Age 20 ($)	Teenage Birth (%)
	(1)	(2)	(3)	(4)
Male	2,235	5.509	1,891	n/a
	(112)	(0.069)	(18)	n/a
	[21,775]	[0.34567]	[24,268]	n/a
Female	2,819	5.828	2,142	−1.028
	(102)	(0.073)	(19)	(0.040)
	[20,889]	[0.42067]	[25,655]	[0.07809]
Non-minority	2,496	5.560	2,911	−0.550
	(172)	(0.098)	(30)	(0.039)
	[31,344]	[0.60147]	[32,288]	[0.01948]
Minority	2,583	5.663	1,624	−1.246
	(80)	(0.058)	(13)	(0.053)
	[17,285]	[0.29627]	[22,031]	[0.10038]
Low Parent Income	2,592	5.209	1,571	−1.210
	(108)	(0.072)	(17)	(0.072)
	[17,606]	[0.27636]	[22,011]	[0.10384]
High Parent Income	2,614	5.951	2,414	−0.834
	(118)	(0.072)	(19)	(0.054)
	[26,688]	[0.49882]	[28,038]	[0.05974]

Notes: Each column reports coefficients from a separate OLS regression. The dependent variable for Column 1 is wage earnings reported on W–2 forms at age 28. The dependent variable for Column 2 is a dummy variable for college attendance at age 20. The dependent variable for Column 3 is our earnings-based college-quality measure. The dependent variable in Column 4 is a dummy variable for having a teenage birth, which we measure using the claiming of a dependent for tax purposes who is fewer than 20 years younger than the individual.

APPENDIX TABLE 6: Cross-Sectional Correlation Between Test Scores and Outcomes in Adulthood by Grade

Dependent Variable:	Earnings at Age 28 ($)	College at Age 20 (%)	College Quality at Age 20 ($)	Earnings at Age 28 ($)	College at Age 20 (%)	College Quality at Age 20 ($)
	No Controls			Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Grade 4	7,618	18.2	5,979	3,252	6.763	2,360
	(77)	(0.053)	(13.8)	(157)	(0.10)	(25.4)
Grade 5	7,640	18.3	6,065	2,498	5.468	1,994
	(62)	(0.052)	(13.6)	(129)	(0.10)	(24.8)
Grade 6	7,395	18.0	5,917	2,103	4.987	1,778
	(63)	(0.057)	(14.7)	(161)	(0.12)	(29.8)
Grade 7	7,790	18.4	5,950	2,308	4.844	1,667
	(65)	(0.060)	(15.5)	(342)	(0.13)	(33.2)
Grade 8	7,591	18.9	6,228	2,133	5.272	1,913
	(55)	(0.055)	(14.1)	(196)	(0.13)	(32.3)
Mean of Dependent Variable	20,867	37.17	24,678	20,867	37.17	24,678

Notes: Each column reports coefficients from a separate OLS regression. Columns 1–3 do not include controls. Columns 4–6 include a full vector of controls. The dependent variable for Columns 1 and 4 is wage earnings reported on W–2 forms at age 28. The dependent variable for Columns 2 and 5 is our earnings-based college-quality measure. The dependent variable for Columns 3 and 6 is our earnings-based index of college quality.

APPENDIX TABLE 7: Robustness Analysis: Clustering and Individual Controls

Dependent Variable:	Score (%)	College at Age 20 (%)	Earnings at Age 28 ($)
	(1)	(2)	(3)
	Panel A: Baseline Analysis Sample		
Baseline estimates	0.861	0.049	1,815
	(0.010)	(0.006)	(727)
95% CI	(0.841, 0.882)	(0.037, 0.062)	(391, 3240)
95% CI using student bootstrap	(0.851, 0.871)	(0.040, 0.056)	(630, 3095)
p value using student bootstrap	<.01	<.01	<.01
	Panel B: Observations with Data on Earnings at Age 28		
	1.157	0.060	1,815
No clustering	(0.016)	(0.010)	(531)
School-cohort	(0.036)	(0.016)	(727)
Two-way student and class	(0.029)	(0.013)	(675)
	Panel C: First Observation for Each Child, by Subject		
Math	0.986	0.040	1,258
No clustering	(0.009)	(0.006)	(780)
School-cohort	(0.017)	(0.007)	(862)
Class	(0.016)	(0.007)	(848)
Reading	1.116	0.061	2544
No clustering	(0.015)	(0.010)	(1320)
School-cohort	(0.025)	(0.012)	(1576)
Class	(0.024)	(0.012)	(1516)
	Panel D: Additional Controls		
Baseline class controls	0.858	0.049	1,696
School-cohort	(0.010)	(0.007)	(797)
Add individual controls	0.856	0.049	1688
School-cohort	(0.010)	(0.007)	(792)
Add school-year effects	0.945	0.026	1942
School-cohort	(0.009)	(0.005)	(669)

Notes: This table replicates the specification from Table 5, Column 1 (first column), Table 7, Column 1 (2nd column), and Table 8, Column 1 (3rd column) using different methods of calculating the standard error and with different control vectors. Panel A reports the results from the baseline specifications, along with the 95%-CI generated from a block-bootstrap at the student level. Panel B reports results on the subsample of observations for whom we have wages at age 28 using both the baseline approach and two-way-clustering by student and classroom (Cameron, Gelbach and Miller, 2011). Panel C takes the first observation per student-subject, eliminating the need to cluster at the individual level, and then reports standard errors from various methods. Finally, Panel D evaluates the sensitivity of the estimates to changes in the control vector. The first and second rows of Panel D use the subsample of observations for which we have data on student-level controls. They compare estimates using two control vectors: the standard classroom-level controls used in Table 2 and other tables vs. the baseline vector plus the student-level controls used to estimate our baseline value-added model (model 1 in Table 3). The third row uses the full analysis sample and includes school-year fixed effects in both the estimation of teacher quality and the outcome regressions.

APPENDIX TABLE 8: Impacts of Test Score: Sensitivity to Trimming

	Percent Trimmed in Upper Tail						Bottom and Top 2%	Jacob and Levitt Proxy
	5%	4%	3%	2%	1%	0%		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
College at Age 20	5.724	5.585	5.258	4.917	4.730	4.022	4.091	6.283
	(0.693)	(0.673)	(0.662)	(0.646)	(0.622)	(0.590)	(0.668)	(0.698)
College Quality at Age 20	1,870	1,848	1,773	1,644	1,560	1,432	1,425	2,033
	(185)	(180)	(177)	(173)	(167)	(160)	(177)	(186)
Earnings at Age 28	2,058	2,080	1,831	1,815	1,581	994	1,719	1,862
	(808)	(776)	(745)	(729)	(709)	(668)	(797)	(821)

Notes: This table replicates results from Columns 1 and 5 of Table 7 (1st and 2nd rows) and Column 1 of Table 8 (3rd row) using different trimming protocols.

APPENDIX TABLE 9: Impacts of Test Score on Earnings by Age

Dependent Variable:	Earnings ($)								
	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Panel A: Full Sample								
Test Score	–211	–322	–211	71	449	558	1,021	1,370	1,815
	(72)	(100)	(136)	(190)	(247)	(319)	(416)	(517)	(729)
Mean Earnings	4,872	6,378	8,398	11,402	13,919	16,071	17,914	19,322	20,353
	Panel B: Low College Attendance								
Test Score	171	165	416	731	1,053	1,405	1,637	1,728	2,073
	(87)	(119)	(159)	(215)	(277)	(343)	(440)	(546)	(785)
Mean Earnings	4,747	6,183	7,785	9,752	11,486	13,064	14,319	15,249	15,967
	Panel C: High College Attendance								
Test Score	–592	–791	–870	–730	–318	–464	448	1,200	2,209
	(110)	(157)	(217)	(317)	(417)	(554)	(717)	(911)	(1,274)
Mean Earnings	5,018	6,609	9,127	13,379	16,869	19,774	22,488	24,718	26,312

Notes: Each column replicates the specification in Column 1 of Table 8, replacing the dependent variable with wages at a given age. In each regression we include observations from all grades. In the second and third regressions, we split the sample based on whether students attend a school with low proportion of students going to college. The mean earnings represents the average of the dependent variable for the observations in the estimation sample. All standard errors are clustered by school-cohort.

APPENDIX TABLE 10: Impacts of Teacher Quality: Instrumental Variables Specifications

Dependent Variable:	Score (SD)	College at Age 20 (%)	College Quality at Age 20 ($)	Earnings at Age 28 ($)	College at Age 20 (%)	College Quality at Age 20 ($)	Earnings at Age 28 ($)
Estimation Method	Reduced Form				Instrumental Variable		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Raw Teacher Quality	0.476	2.526	786	871			
	(0.006)	(0.349)	(93)	(392)			
Score					5.29	1,647	1,513
					(0.72)	(191)	(673)
Observations	3,721,120	3,095,822	3,095,822	368,427	3,089,442	3,089,442	368,427
Mean of Dependent Variable	0.162	37.8	24,713	20,912	37.8	24,713	20,912

Notes: This table reproduces the main specifications in Table 7 (Columns 1 and 5) and Table 8 (Column 1) using unshrunken teacher quality estimates, which are precision-weighted averages of student score residuals in other classes taught by a given teacher. Columns 1 through 4 regress the outcome on unshrunken teacher quality with controls. Columns 5 through 7 instrument for score with unshrunken teacher quality including controls. All regressions cluster standard errors at the school X-cohort level.