NBER WORKING PAPER SERIES

HOW DOES YOUR KINDERGARTEN CLASSROOM AFFECT YOUR EARNINGS? EVIDENCE FROM PROJECT STAR

Raj Chetty
John N. Friedman
Nathaniel Hilger
Emmanuel Saez
Diane Whitmore Schanzenbach
Danny Yagan

Working Paper 16381
http://www.nber.org/papers/w16381

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2010

We thank Lisa Barrow, David Card, Elizabeth Cascio, Janet Currie, Jeremy Finn, Edward Glaeser, Bryan Graham, James Heckman, Caroline Hoxby, Thomas Kane, Lawrence Katz, Alan Krueger, Derek Neal, Jonah Rockoff, Douglas Staiger, and numerous seminar participants for helpful discussions and comments. We thank Helen Bain and Jayne Zaharias at HEROS for access to the Project STAR data. The tax data were accessed through contract TIRNO-09-R-00007 with the Statistics of Income (SOI) Division at the US Internal Revenue Service. We thank the SOI staff, and in particular Nicholas Greenia, for their invaluable help and guidance in this process. Gregory Bruich, Jane Choi, Jessica Laird, Keli Liu, Laszlo Sandor, and Patrick Turley provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard, the Center for Equitable Growth at UC Berkeley, and the National Science Foundation is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR
Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan
NBER Working Paper No. 16381
September 2010, Revised October 2010
JEL No. H0,J0

ABSTRACT

In Project STAR, 11,571 students in Tennessee and their teachers were randomly assigned to different classrooms within their schools from kindergarten to third grade. This paper evaluates the long-term impacts of STAR using administrative records. We obtain five results. First, kindergarten test scores are highly correlated with outcomes such as earnings at age 27, college attendance, home ownership, and retirement savings. Second, students in small classes are significantly more likely to attend college, attend a higher-ranked college, and perform better on a variety of other outcomes. Class size does not have a significant effect on earnings at age 27, but this effect is imprecisely estimated. Third, students who had a more experienced teacher in kindergarten have higher earnings. Fourth, an analysis of variance reveals significant kindergarten class effects on earnings. Higher kindergarten class quality — as measured by classmates' end-of- class test scores — increases earnings, college attendance rates, and other outcomes. Finally, the effects of kindergarten class quality fade out on test scores in later grades but gains in non-cognitive measures persist. We conclude that early childhood education has substantial long-term impacts, potentially through non-cognitive channels. Our analysis suggests that improving the quality of schools in disadvantaged areas may reduce poverty and raise earnings and tax revenue in the long run.

Raj Chetty
Department of Economics
Harvard University
1805 Cambridge St.
Cambridge, MA 02138
and NBER
chetty@fas.harvard.edu

John N. Friedman
Harvard Kennedy School
Taubman 356
79 JFK St.
Cambridge, MA 02138
and NBER
jfriedman@post.harvard.edu

Nathaniel Hilger
Department of Economics
Harvard University
1805 Cambridge St.
Cambridge, MA 02138
nhilger@fas.harvard.edu

Emmanuel Saez
Department of Economics
University of California, Berkeley
549 Evans Hall #3880
Berkeley, CA 94720
and NBER
saez@econ.berkeley.edu

Diane Whitmore Schanzenbach
School of Education and Social Policy
Northwestern University
Annenberg Hall, Room 205
2120 Campus Drive
Evanston, IL 60208
and NBER
dws@northwestern.edu

Danny Yagan
Department of Economics
Littauer 200, North Yard
Harvard University
Cambridge, MA 02138
yagan@fas.harvard.edu

I Introduction

What are the long-term impacts of early childhood education? Evidence on this important policy question remains scarce because of a lack of data linking childhood education and adult outcomes. This paper analyzes the long-term impacts of Project STAR, one of the most widely studied education experiments in the United States. The Student/Teacher Achievement Ratio (STAR) experiment randomly assigned one cohort of 11,571 students and their teachers to different classrooms within their schools in grades K-3. Some students were assigned to small classes (15 students on average) in grades K-3, while others were assigned to large classes (22 students on average). The experiment was implemented across 79 schools in Tennessee from 1985-89. Numerous studies have used the STAR experiment to show that class size, teacher quality, and peers have significant causal impacts on test scores.[1] Whether these gains in achievement on standardized tests translate into improvements in adult outcomes such as earnings remains an open question.

We link the original STAR data to administrative data from tax returns, allowing us to follow 95% of the STAR participants into adulthood.[2]

These linked data allow us to measure the impacts of early childhood education on outcomes ranging from college attendance and earnings to retirement savings, home ownership, and marriage. The dataset also contains information on the characteristics of each student's parents. The linked dataset overcomes two concerns about prior research on STAR: attrition and insufficient baseline data to fully evaluate the randomization protocol. Attrition is not a serious concern because we are able to track virtually all students, and there are no significant differences in match rates across classrooms. To confirm the validity of the randomization protocol, we show that several parent characteristics that predict student outcomes – such as the age at which the child was born and parent's income – are all balanced across classrooms within schools.

We begin our empirical analysis by documenting the strong correlation between kindergarten test scores and adult outcomes in the cross section. A one percentile increase in end-of-kindergarten (KG) test scores is associated with a $132 increase in wage earnings at age 27 (relative to the mean

[1] On class size, see e.g. Finn and Achilles (1990), Word et al. (1990), and Krueger (1999). On teacher effects, see e.g. Nye, Konstantopoulos, and Hedges (2004) and Dee (2004). On peer effects, see e.g. Cascio and Schanzenbach (2007), Graham (2008), and Sojourner (2009). Schanzenbach (2006) provides a comprehensive review of the STAR literature. Outside the STAR literature, numerous studies have used quasi-experimental methods to estimate the effects of class size and other educational inputs on test scores (e.g. Angrist and Lavy 1999, Hoxby 2000, Rivkin, Hanushek, and Kain 2005).

[2] The data for this project were analyzed through a program developed by the Statistics of Income (SOI) Division at the U.S. Internal Revenue Service to support research into the effects of tax policy on economic and social outcomes and improve the administration of the tax system.

$15,912) in the raw data and a $94 increase controlling for parental characteristics. Equivalently, a one standard deviation (SD) increase in test scores is associated with an 18% increase in earnings conditional on parent characteristics.[3] A one percentile increase in KG test scores is also associated with a 0.53 percentage point increase in the probability of attending college by age 27. We construct a new index of college quality based on the mean earnings of students who attended each college. By this measure, as well as other standard college rankings, increases in KG test scores are strongly associated with attending a better college. Several other adult outcomes – such as home ownership, 401(k) savings, mobility rates, neighborhood quality, and marital status – are also all highly correlated with kindergarten test scores. These strong correlations motivate the main question of the paper: do interventions that raise test scores – such as smaller classes and better teachers – cause analogous improvements in adult outcomes?

Our analysis of the experimental impacts consists of two empirical strategies. First, we study the impacts of observable classroom characteristics. We analyze class size effects using the same intent-to-treat specifications as in Krueger (1999), who shows that students assigned to small classes score approximately 4.8 percentile points (0.2 standard deviations) higher than students in large classes on tests in kindergarten. We find that students assigned to small classes are 1.8 percentage points more likely to be enrolled in college at age 20, a significant improvement relative to the mean college attendance rate of 26.4% at age 20 in the sample. Students in small classes also exhibit statistically significant improvements on a summary index of the other outcomes we examine (home ownership, 401(k) savings, mobility rates, percent college graduate in ZIP code, and marital status). We do not find significant differences in earnings at age 27 between students who were in small and large classes, but these earnings impacts are imprecisely estimated. In particular, we cannot reject the hypothesis that small classes generate an earnings gain commensurate to that implied by the cross sectional correlation ($4.8 \times \$94 = \$451$).

We study variation across classrooms along other observable dimensions, such as teacher and peer characteristics, using a similar approach. Prior studies (e.g. Krueger 1999) have shown that STAR students with more experienced teachers score higher on tests. We find similar impacts on earnings. Students randomly assigned to a KG teacher with more than 10 years of experience earn an extra $\$1,093$ (6.9% of mean income) on average at age 27 relative to students with less experienced teachers.[4] Consistent with earlier research on teacher effects on test scores, we are

[3] These cross-sectional estimates are consistent with those obtained by Currie and Thomas (2001) using the British National Child Development Survey and Currie (2010) using the National Longitudinal Survey of Youth.

[4] Because teacher experience is likely to be correlated with many other unobserved attributes – such as attachment

unable to identify other observable teacher characteristics (e.g. degrees) that have a significant impact on adult outcomes. We test whether observable peer characteristics have long-term impacts by regressing earnings on the fraction of low-income, female, and black peers in KG. These peer impacts are not significant and are very imprecisely estimated. This lack of power in detecting peer effects is not surprising given that the experimental design randomized students across classes, generating relatively little variation in peer characteristics across classes.

Because we have few measures of observable classroom characteristics, we turn to a second empirical strategy that captures both observed and unobserved aspects of classrooms. We use an analysis of variance approach analogous to that in the teacher effects literature (Hanushek 1971, Murnane 1975, Rockoff 2004, Rivkin, Hanushek, and Kain 2005, Aaronson, Barrow, and Sander 2007). Because we observe each teacher only once in our data, we can only estimate "class effects" – the combined effect of teachers, peers, and any class-level shock such as noise outside the classroom – by exploiting random assignment to KG classrooms of both students and teachers. Intuitively, we test whether earnings vary across KG classes by more than what would be predicted by random variation in student abilities. An F test rejects the null hypothesis that KG classroom assignment has no effect on earnings with $p < 0.05$. The standard deviation of class effects on annual earnings is \$1,520 (9.6% of mean income). This result highlights the stakes at play in early childhood education: with a discount rate of 3%, a kindergarten class of 20 students that is one standard deviation better in quality generates nearly \$782,000 more in present value earnings on average.

The analysis of variance shows that kindergarten classroom assignment has significant impacts on earnings, but does not tell us whether educational interventions that improve scores also generate earnings gains. To analyze this question, we proxy for each student's KG "class quality" by the average test scores of his classmates at the end of kindergarten. End-of-class peer test scores are an omnibus measure of class quality because they capture both ex-ante variation in peer abilities and the effects of the teacher on students. Using this measure, we find that kindergarten class quality has significant impacts on both test scores and earnings. Students randomly assigned to a class that is one standard deviation higher in quality score 6.27 percentile points higher on end-of-year tests and earn \$483 (3.0%) more at age 27. This implies an earnings gain of \$77 (0.5%) per percentile improvement in test scores.

The earnings gains remain large and significant when conditioning on observable classroom char-

to the teaching profession – we cannot conclude that increasing teacher experience would improve student outcomes. This evidence simply establishes that a student's KG teacher has causal impacts on his or her earnings as an adult.

acteristics, confirming that much of the impact of classrooms on adult outcomes is driven by features of classroom environments that are not observed in our data. KG classroom assignments have significant impacts on adult outcomes even for students in large classes, who were re-randomized after kindergarten. This shows that a single year of high quality early childhood classroom education (rather than a string of good classrooms over multiple years) has long-lasting impacts. Students assigned to higher quality classes also show improvements on all the other adult outcomes: they are more likely to attend college, enroll in higher quality colleges, and exhibit improvement in the summary index of other outcomes. The effects of class quality are widespread: blacks and whites, males and females, and free-lunch eligible and ineligible children all exhibit improvements across most of the adult outcomes when assigned to higher quality classes.

Although we emphasize the effects of kindergarten, the class quality impacts are similar for students who were randomized into classes in grades 1-3 upon entry into the STAR project. Hence, the findings of this paper should be viewed as evidence on the long-term impacts of early childhood education rather than kindergarten in particular. For new entrants, the key determinant of adult outcomes is end-of-year peer test scores rather than peer test scores in the previous year. We also find that end-of-year scores matter more than predicted peer scores based on observable characteristics. Both of these findings suggest that teachers and class-level shocks (which include interactions between peers and teachers) have greater impacts on outcomes in the STAR sample than ex-ante variation in peer abilities.[5]

The impacts of early childhood class assignment on outcomes more than twenty years later is surprising in view of evidence that the impacts "fade out" on test scores. The effects of small class size on test scores falls to 1-2 percentile points (0.05 standard deviations) by grade 8 (Krueger and Whitmore 2001). We find similar fade-out on the effects of class quality on test scores by 8th grade. Why do the impacts of early childhood education fade-out on test scores but re-emerge in adulthood? We find some suggestive evidence that one key channel may be non-cognitive skills. In particular, KG class quality has significant impacts on non-cognitive measures in 4th and 8th grade such as effort, initiative, and disruptive behavior. These non-cognitive measures are highly correlated with earnings even conditional on test scores. Persistent gains in non-cognitive skills provide an intuitive explanation of our findings. Kindergarten classroom environments that teach students to do well on KG math and reading tests may build skills such

[5]This does not imply that peers do not matter. Our point is simply that most of the variation in quality across classrooms in the STAR experiment is not driven by ex-ante peer characteristics.

as focus and discipline. These non-cognitive skills might not be well measured in tests of cognitive ability, but have substantial returns in the labor market and improve adult outcomes. While this evidence is not conclusive, it highlights the value of further empirical research on non-cognitive skills.

In addition to the extensive literature on the impacts of STAR on test scores, our study builds on and contributes to a recent literature investigating selected long-term impacts of class size in the STAR experiment. These studies have shown that students assigned to small classes are more likely to complete high school (Finn, Gerber, and Boyd-Zaharias 2005) and take the SAT or ACT college entrance exams (Krueger and Whitmore 2001).[6] Students in small classes are less likely to be arrested for crime and white females in small classes have lower teen birth rates (Krueger and Whitmore 2001). Most recently, Muennig et al. (2010) report that students in small classes have higher mortality rates, a finding that we do not obtain in our data as we discuss below. Our contribution to this literature is twofold. First, we provide a unified evaluation of impacts on several outcomes, including the first analysis of earnings. Second, we examine the impacts of teachers and peers and in addition to class size.

Our results also complement the findings of studies on the long-term impacts of other early childhood interventions (reviewed in Almond and Currie 2010). For example, the Perry preschool program randomized 123 children into a control group and an intensive pre-school treatment group. Schweinhart, Barnes, and Weikhart (1993), Schweinhart et al. (2005), and Heckman et al. (2010a, 2010b) show that the Perry preschool program had extremely large impacts on earnings and other adult outcomes despite relatively rapid fade-out of test score impacts. Heckman et al. (2010c) show that the Perry intervention also improved non-cognitive skills and argue that this mechanism accounts for much of the long-term impact. Campbell et al. (2002) show that the Abecedarian project, which randomized 111 children into intensive early childhood programs from infancy to age 5, led to lasting improvements in education and other outcomes in early adulthood. At a larger scale, several studies have shown that the Head Start program leads to improvement in a variety of adult outcomes despite fade-out on test scores (e.g., Currie and Thomas 1995, Garces, Thomas, and Currie 2002, Ludwig and Miller 2007, Deming 2009). The results reported here are the first experimental evidence on the long-term impacts of a scalable intervention in a large sample with minimal attrition. In particular, we show that a better classroom environment from ages 5-8

[6] Dynarski, Hyman, and Schanzenbach (2010) are currently investigating the impact of class size on college attendance using data from the National Student Clearinghouse.

can have substantial long-term benefits even without intervention at earlier ages. This result is consistent with the findings of Card and Krueger (1992), who show that better educational inputs have substantial long-term payoffs using state-by-cohort variation.

Finally, our research has important implications for tax policy and tax administration. Some of the payoffs from better education will accrue to the government via increased tax payments and reduced dependence on welfare programs, underscoring the importance of early childhood education for tax and transfer policy. Therefore, our research contributes and develops a new method for estimating the effect of educational public policies on tax revenue. In terms of tax administration, our research shows that the administration of federal tax credits for higher education provides comprehensive information on college enrollment in the United States. Therefore, as our study demonstrates, the federal tax system and the tax data it generates can play an important role in the evaluation of US education policy.

The paper is organized as follows. In section II, we review the STAR experimental design, summarize the linked dataset, and address potential threats to the validity of the experiment. Section III documents the cross-sectional correlation between kindergarten test scores and adult outcomes. Section IV analyzes the impacts of observable characteristics of classrooms – size, teacher characteristics, and peer characteristics – on adult outcomes. In Section V, we study class effects more broadly, incorporating unobservable aspects of class quality. Section VI documents the fade-out and re-emergence effects and the potential role of non-cognitive skills in explaining this pattern. In the concluding section, we use the empirical estimates to calculate the benefits of class size reductions and improvements in teacher quality.

II Experimental Design and Data

II.A Background on Project STAR

Word et al. (1990), Krueger (1999), and Finn et al. (2007) provide a comprehensive summary of Project STAR; here, we briefly review the features of the STAR experiment most relevant for our analysis. The STAR experiment was conducted in 79 schools across the state of Tennessee for 4 years. Participating schools were drawn from areas across the state. The program oversampled lower-income schools, and thus the STAR sample exhibits lower socioeconomic characteristics than the state of Tennessee and the US population as a whole.

In the 1985-86 school year, 6,323 kindergarten students in participating schools were randomly assigned to a small (target size 13-17 students) or regular-sized (20-25 students) class within their

schools.[7] Students were intended to remain in the same class type (small vs. large) through 3rd grade, at which point all students would return to regular classes for 4th grade and subsequent years. As the initial cohort of kindergarten students advanced across grade levels, there was substantial attrition because students who moved away from a participating school or were retained in grade no longer received treatment. In addition, because kindergarten was not mandatory and due to normal residential mobility, many children entered the participating schools after kindergarten. A total of 5,248 students entered the participating schools in grades 1-3. These new entrants were randomly assigned to classrooms upon entry into one of the participating schools. As a result, the randomization pool is school-by-entry wave, and we include school-by-entry wave fixed effects in all experimental analyses below.

Upon entry into one of the 79 schools, the study design randomly assigned students not only to class type (small vs. large) but also to a classroom within each type (if there were multiple classrooms per type, as was the case in 50 of the 79 schools). Teachers were also randomly assigned to classrooms. Unfortunately, the exact protocol of randomization into specific classrooms was not clearly documented in any of the official STAR reports, where the emphasis was instead the random assignment into class type rather than classroom (Word et al. 1990). We present statistical evidence confirming that both students and teachers indeed appear to be randomly assigned directly to classrooms as the original designers attest.[8] After kindergarten, it was left to the school's discretion whether to remix the previously enrolled students across classrooms within class type or to maintain the original classroom assignments.

As in any field experiment, there were some deviations from the experimental protocol. In particular, some students moved from large to small classes and vice versa. To account for such potentially non-random sorting, we adopt the standard approach taken in the literature and assign treatment status based on initial random assignment (intent-to-treat).

In each year, students were administered the grade-appropriate Stanford Achievement Test, a multiple choice test that measures performance in math and reading. Following Krueger (1999), we standardize the math and reading scale scores in each grade by computing the scale score's

[7]There was also a third treatment group: regular sized class with a full-time teacher's aide. This was a relatively minor intervention, since all regular classes were already assigned a 1/3 time teacher's aide. Prior studies of STAR find no impact of a full-time teacher's aide on test scores. We follow the convention in the literature and group the regular and regular plus aide class treatments together. The results reported below are unaffected by including an indicator for the aide treatment, as we find no impacts of that treatment on any adult outcome.

[8]While studies that analyze class size effects require only randomization within class type, studies of teacher and peer effects (e.g. Dee 2004, Nye, Konstantopoulos, and Hedges 2004, Cascio and Schanzenbach 2007, Graham 2008) exploit randomization of students into classrooms.

corresponding percentile rank in the distribution for students in large classes. We then assign the appropriate percentile rank to students in small classes, and take the average across math and reading percentile ranks. Note that this percentile measure is a ranking of students *within* the STAR sample. We were unable to obtain the information needed to benchmark the K-3 scores to a national population. However, we can compare the distribution on the Stanford Achievement Tests given to students in 4th through 8th grades to a nationally representative sample. The nationally normed standard deviation on these tests in grades 4-8 is only 1.02 times the corresponding standard deviation within our sample.[9] Hence, the within-sample percentile test score measures used below can also be approximately interpreted as percentiles in the national distribution.

II.B Variable Definitions and Summary Statistics

We measure adult outcomes of Project STAR participants using administrative data from United States tax records. Selected variables from tax records were accessed through a research contract with the Statistics of Income (SOI) division of the Internal Revenue Service. 95.0% of STAR records were linked to the tax data using an algorithm based on standard identifiers (SSN, date of birth, gender, and names) that is described in Appendix A.[10]

We obtain data on students and their parents from federal tax forms such at 1040's. Information from 1040's is available from 1996-2008. Approximately 10% of adults do not file individual income tax returns in a given year. We use third-party reports to obtain information such as earnings (form W-2), unemployment benefits received (form 1099), and college attendance (form 1098-T) for these non-filers. Data from these third-party reports are available since 1999. In all cases, 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 $t+1$. The analysis dataset combines selected variables from individual tax returns, third party reports, and information from the STAR database, with individual identifiers removed to protect confidentiality.

We now describe how each of the adult outcome measures and control variables used in the empirical analysis is constructed. Table 1 reports summary statistics for these variables for the STAR sample as well as a random 0.25% sample of the US population born in the same years (1979-80).

[9] Compared to the overall K-3 sample, the sample of STAR students for whom follow-up test scores are available is slightly compressed. The ratio between the two standard deviations is approximately 1.02.

[10] The matching algorithm was sufficiently precise that it uncovered 28 cases in the original STAR dataset that were a single split observation or duplicate records. After consolidating these records, we are left with 11,571 students.

Earnings. The individual earnings data come from W-2 forms, yielding information on earnings for both filers and non-filers.[11] We define earnings in each year as the sum of earnings on all W-2 forms filed on an individual's behalf. We always express all monetary variables in 2009 dollars, adjusting for inflation using the Consumer Price Index. We winsorize earnings in each year at $100,000 to eliminate outliers; fewer than 1% of individuals in the STAR sample report earnings above $100,000 in a given year. To increase precision, we typically use average (inflation indexed) earnings from year 2005 to 2007 as an outcome measure. The mean individual earnings for the STAR sample in 2005-2007 (when the STAR students are 25-27 years old) is $15,912. This earnings measure includes zeros for the 13.9% of STAR students who report no income between 2005-07. The mean level of earnings in the STAR sample is lower than in the same cohort in the US population, as expected given that Project STAR targeted more disadvantaged schools.

College Attendance. Higher education institutions that are eligible for federal financial aid (i.e., Title IV institutions) are required to file 1098-T forms that report tuition payments or scholarships received for every student.[12] Title IV institutions include all colleges and universities as well as vocational schools and other postsecondary institutions. Comparisons to other data sources indicate that 1098-T forms accurately capture college enrollment in the U.S.[13] We have data on college attendance from 1098-T forms for all students in our sample since 1999, when the STAR students were 19 years old. We define college attendance as an indicator for having one or more 1098-T forms filed on one's behalf in a given year. In the STAR sample, 26.4% of students are enrolled in college at age 20 (year 2000). 45.5% of students are enrolled in college at some point between 1999 and 2007, compared with 57.1% in the same cohort of the US population. Because the data are based purely on tuition payments, we have no information about college completion or degree attainment in these data.

College Quality. Using the institutional identifiers on the 1098-T forms, we construct an earnings-based index of college quality as follows. First, using the 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

[11] We obtain similar results using household adjusted gross income reported on individual tax returns. We focus on the W-2 measure because it is 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 reflect self-employment income.

[12] These forms are used to administer the Hope and Lifetime Learning education tax credits created by the Taxpayer Relief Act of 1997.

[13] In 2009, 27.4 million 1098-T forms were issued (Internal Revenue Service, 2010). According to the Current Population Survey (US Census Bureau, 2010, Tables 5 and 6), in October 2008, there were 22.6 million students in the U.S. (13.2 million full time, 5.4 million part-time, and 4 million vocational). As an individual can be a student at some point during the year but not in October, the number of 1098-T forms for the calendar year should be higher than the number of students as of October.

by the higher education institution they attended in 1999. Individuals who attended more than one institution in 1999 are counted as students at all institutions they attended. Those not attending a higher education institution in 1999 are pooled together in a separate "no college" category. Next, we compute average earnings of the students in 2007 when they are aged 27 by their educational institution in 1999. This earnings-based index of college quality is highly correlated with the US News ranking of the best 125 colleges and universities: the correlation coefficient of our measure and the log US news rank is 0.75. The advantages of our index are that it covers all higher education institutions in the U.S. and provides a simple cardinal metric for college quality. Among colleges attended by STAR students, the average value of our earnings index is $35,080 for four-year colleges and $26,920 for two-year colleges.[14] For students who did not attend college, the imputed mean wage is $16,475.

Other Outcomes. We identify spouses using information from 1040 forms. For individuals who file tax returns, we define an indicator for marriage based on whether the tax return is filed jointly. We code non-filers as single because most non-filers in the U.S. who are not receiving Social Security benefits are single (Cilke 1998, Table 1). We define a measure of ever being married by age 27 as an indicator for ever filing a joint tax return in any year between 1999 and 2007. By this measure, 43.2% of individuals are married at some point before age 27.

We measure retirement savings using contributions to 401(k) accounts reported on W-2 forms from 1999-2007. 28.2% of individuals in the sample make a 401(k) contribution at some point during this period. We measure home ownership using data from the 1098 form, a third party report filed by lenders to report mortgage interest payments. We include the few individuals who report a mortgage deduction on their 1040 forms but do not have 1098's as homeowners. We define any individual who has a mortgage interest deduction at any point between 1999 and 2007 as a homeowner. Note that this measure of home ownership does not cover individuals who own homes without a mortgage, which is rare among individuals younger than 27. By our measure, 30.8% of individuals own a home by age 27. 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 did not file and has no W-2 in a given year, we impute current ZIP code as the last observed ZIP code. We define a measure of cross-state mobility by an indicator for whether the individual ever lived outside Tennessee between 1999 and 2007.

[14] For the small fraction of STAR 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.

27.5% of STAR students have lived outside Tennessee at some point between age 19 and 27. We construct a measure of neighborhood quality using data on the percentage of college graduates in the individual's 2007 ZIP code from the 2000 Census. On average, STAR students lived in neighborhoods with 17.6% college graduates in 2007.

We observe dates of birth and death until the end of 2009 as recorded by the Social Security Administration. We define each STAR participant's age at kindergarten entry as the student's age (in days divided by 365.25) as of September 1, 1985. Virtually all students in STAR were born in 1979-80. To simplify the exposition, in the text we say that the cohort of STAR children are aged a in year $1980 + a$ (e.g. STAR children are 27 in 2007). Approximately 1.7% of the STAR sample is deceased by 2009.[15]

Parent Characteristics. We link STAR children to their parents by finding the earliest 1040 form from 1996-2008 on which the STAR children were claimed as dependents. Most matches were found on 1040 forms for tax year 1996, when the STAR children were 16. We identify parents for 86% of the STAR students in our linked dataset. The remaining students are likely to have parents who did not 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.

We define parental household income as average Adjusted Gross Income (winsorized at $252,000, the 99th percentile) from 1996-1998, when the children are 16-18 years old. For parents who do not file, we define household income as zero.[16] For divorced parents, this income measure captures the total resources available to the household claiming the child as a dependent (including any alimony payments), rather than the sum of the individual incomes of the two parents. By this measure, mean parent income is $48,010 (in 2009 dollars) for STAR students whom we are able to link to parents. We define marital status, home ownership, and 401(k) saving as indicators for whether the parent who claims the STAR 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

[15] Aggregate death rates in the Social Security Death Master file are relatively close to official US vital statistics in recent years. As 95% of STAR students are matched to the tax data and have a valid Social Security Number, we believe that deaths are recorded accurately in our sample.

[16] Alternative definitions of income for non-filers – such as income reported on W-2's starting in 1999 – yield very similar results to those reported below.

using the age of the filer who claimed the child, who is typically the mother but is sometimes the father or another relative.[17] By this measure, mothers are on average 25.0 years old when they give birth to a child in the STAR sample. When a child cannot be matched to a parent, we define all parental characteristics as zero and always include a dummy for missing parents.

Background Variables from STAR. In addition to classroom assignment and test score variables, we use some demographic information from the STAR database in our analysis. This includes gender, race (an indicator for being black), and whether the student ever received free or reduced price lunch during the experiment. 36% of the STAR sample are black and 60% are eligible for free or reduced-price lunches. In our linked dataset, race is missing for 31 observations and free lunch status is missing for 146 observations. Finally, we use data on teacher characteristics – experience, race, and highest degree – from the STAR database. The average kindergartner has a teacher with 9.3 years of experience. 16.2% of kindergarten students have a black teacher, and 34.8% have a teacher with a master's degree or higher.

II.C Validity of the Experimental Design

The validity of the causal inferences that follow rests on two assumptions: successful randomization of students into classrooms and no differences in attrition (match rates) across classrooms. We now evaluate each of these issues.

Randomization into Classrooms. To evaluate whether the randomization protocol was implemented as designed, we test for balance in pre-determined covariates across classrooms. The original STAR dataset contains only a few pre-determined variables: age, gender, race, and free-lunch status. Although the data are balanced on these characteristics, some skepticism naturally has remained because of the coarseness of the variables (Hanushek 2003).

The tax data allow us to improve upon the prior evidence on the validity of randomization by investigating a wider variety of family background characteristics. In particular, we check for balance in the following five parental characteristics: household income, 401(k) savings, home ownership, marital status, and mother's age at child's birth. Although most of these characteristics are not measured prior to random assignment in 1985, they are measured prior to the STAR cohort's expected graduation from high school and are unlikely to be impacted by the child's classroom assignment in grades K-3. We first establish that these parental characteristics are in fact strong

[17] 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 a birth date recorded in the SSA file.

predictors of student outcomes. In Column 1 of Table 2a, we regress the child's first available test score on the five parent characteristics, the student's age, gender, race, and free-lunch status, and school by entry wave fixed effects.[18] We also include indicators for missing data on certain variables (parents' characteristics, mothers's age, student's free lunch status, and student's race). The student and parent demographic characteristics are highly significant predictors of test scores.

Having identified a set of pre-determined characteristics that predict children's performance in school, we test for balance in these covariates across classrooms. We first evaluate randomization into the small class treatment by regressing an indicator for being assigned to a small class upon entry on the same variables as in column 1. As shown in column 2 of Table 2, none of the demographic characteristics predict the likelihood that a child is assigned to a small class. An F test for the joint significance of all the pre-determined demographic variables is insignificant ($p = 0.26$), showing that students in small and large classes have similar demographic characteristics.

Next, we evaluate whether students were randomly assigned into classrooms within small or large class types. If students were randomly assigned to classrooms, then conditional on school fixed effects, classroom indicator variables should not predict any pre-determined characteristics of the students. Column 6 of Table 2 reports p values from F tests for the significance of kindergarten classroom dummies in regressions of each pre-determined characteristic on class and school fixed effects. None of the F tests is significant, showing that the parental and child characteristics are balanced across classrooms.

Columns 3-5 of Table 2 evaluate the random assignment of teachers to classes by regressing teacher characteristics – experience, bachelor's degree, and race – on the same student and parent characteristics. Again, none of the pre-determined variables predict the type of teacher a student is assigned, consistent with random assignment of teachers to classrooms.

Selective Attrition. Another threat to the experimental design is differential attrition across classrooms. Attrition is a much less serious concern in the present study than in past evaluations of STAR (Hanushek 2003) because we are able to locate 95% of the students in the tax data. Nevertheless, we investigate whether the likelihood of being matched to the tax data varies by classroom assignment within schools. In columns 1 and 2 of Table 3, we test whether the match rate varies significantly with class size by regressing an indicator for being matched on the small

[18]The first available test score depends upon the point at which the student enters a participating school. The first available score is the kindergarten score for 55% of the sample, first grade score for 23%, second grade score for 12%, and third grade score for the remaining 11%. If a student's test score from her initial year of enrollment is missing, we do not take a later score but rather leave the observation missing.

class dummy. Column 1 includes no controls other than school by entry wave fixed effects. It shows that eliminating the between-school variation, the match rate in small and large classes differs by less than 0.02 percentage points. Column 2 shows that controlling for the full set of demographic characteristics used in Table 2a does not uncover any significant difference in the match rate across class types. The p values reported at the bottom of columns 1 and 2 are for F tests of the significance of classroom indicators in predicting match rates in regression specifications analogous to those in Column 6 of Table 2. The p values are approximately 0.9, showing that there are no significant differences in match rates across classrooms within schools.

Another potential source of attrition from the sample is through death. Columns 3 and 4 replicate Columns 1 and 2 of Table 3, replacing the dependent variable in the regressions with an indicator for death before January 1, 2010. We find no evidence that mortality rates vary with class size or across classrooms. The difference in death rates between small and large classes is less than 0.01 percentage points. These findings are inconsistent with recent results reported by Muennig et al. (2010), who find that students in small classes are slightly more likely to die using data from the National Death Index. The discrepancy between the findings might be due to differences in match quality.

The tests in Tables 2 and 3 apply to the pooled sample of STAR students. We replicated these randomization and selective attrition tests for students who entered the participating schools after kindergarten and found that covariates and match rates are balanced across classrooms in each entry wave.

III Test Scores and Adult Outcomes in the Cross-Section

We begin by documenting the correlations between kindergarten test scores and adult outcomes in the cross-section. We estimate models similar to those in Murnane, Willett, and Levy (1995) and Neal and Johnson (1996), who report correlations between test scores in high school and earnings. This descriptive analysis provides a benchmark to assess the impacts of the randomized interventions that improve test scores. Note that only 5,621 of the STAR participants who enter in KG both have a KG test score and are matched to the tax data. We use only these observations in the descriptive analysis in this section for simplicity.

We first examine the correlation between KG test scores and mean earnings from ages 25-27. Although individuals' earnings trajectories remain quite steep at age 27, earnings levels from ages

25-27 are highly correlated with earnings at later ages (Haider and Solon 2006).[19] Figure 1a shows that there is a strong association between KG test scores and mean earnings from age 25-27. To construct this figure, we bin individuals into twenty equal-sized bins and plot mean earnings in each bin. A one percentile point increase in KG test score is associated with a $132 (0.83%) increase in earnings twenty years later, as shown in column 1 of Table 4a. The correlation between KG test score percentiles and earnings remains significant even in the tails of the distribution of test scores, showing the considerable predictive power of tests administered in early childhood. However, KG test scores explain only a small share of the variation in adult earnings: the R^2 of the regression of earnings on scores is 5%.

The remaining specifications of Table 4a probe how changes in controls and specification affect this relationship. In column 2, we include classroom fixed effects, in order to isolate the non-experimental variation that arises purely from differences across students in the cross-section. This yields a slightly larger coefficient of $143 per percentile. Column 3 controls for a vector of parent and student demographic characteristics. The parent characteristics are a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, and indicators for parent's 401(k) savings and home ownership. The student characteristics are gender, race, age at KG entry, and free lunch status. We code all parental characteristics as 0 for students whose parents are missing, and include a dummy for missing parents as a control. We also include indicators for missing data on certain variables (mothers's age, student's free lunch status, and student's race) and code these variables as zero when missing. We use this vector of demographic characteristics in most specifications below. When the demographic controls are included, the coefficient on kindergarten percentile scores falls to $94, showing that part of the raw correlation in Figure 1 is driven by these characteristics.

In column 4, we regress log wage earnings on test scores in standard deviation units, as in Neal and Johnson (1996).[20] A one standard deviation increase in KG test score is associated with an 18.0% increase in earnings from ages 25-27. This finding is quite similar to that in Neal and Johnson (1996) and Hanushek and Zhang (2009), who find that a 1 standard deviation increase in test scores measured between ages 16-18 increases wage earnings measured at ages 26-29 by

[19] We replicate this finding in our data. Appendix Table 1 reports the correlation between earnings at age x and age $x+6$ (the maximal span over which we can observe earnings in our data) from age 18-50 in the U.S. population. The correlation of earnings across years is maximized around age 40 at 0.8, but even at age 26 the correlation is 0.65.

[20] This specification omits the 13.9% of individuals who have zero earnings. We use level of earnings as the dependent variable in our primary specifications to avoid selection problems from omitting individuals who do not work. We use the percentile test score (rather than the score measured in standard deviation units) because earnings appear to vary linearly with percentiles rather than the raw scores, as shown in Figure 1a.

20%.[21] The similarity between the coefficients obtained for kindergarten and high school test scores suggests that kindergarten scores contain much of the same information as later test scores. In column 5 of Table 4a, we demonstrate this result more directly by including both 8th grade scores (the last point at which data from standardized tests are available for most students in the STAR sample) and KG scores in the regression. The inclusion of 8th grade scores raises the R^2 only slightly, from 16.6% to 18.1%.

Finally, in column 6 of Table 4a, we compare the relative importance of parent characteristics and cognitive ability as measured by test scores. We calculate the parent's income percentile rank using the tax data for the U.S. population. Conditional on test scores, a one percentile point increase in parental income is associated with approximately a $158 increase in earnings, suggesting that parental background affects earnings as much as or more than cognitive ability in the cross section.

Figures 1b and 1c show the cross-sectional relationship between kindergarten test score and college outcomes. Columns 1-3 in the first row of Table 4b show the corresponding regression coefficients, conditional on the full vector of controls used in column 3 of Table 4a. Figure 1b shows that KG test scores are highly predictive of college attendance. Conditional on demographic characteristics, a one percentile point increase in KG test score is associated with a 0.40 percentage point increase in the probability of attending college at age 20 and a 0.53 percentage point increase in the probability of attending college at some point before age 27. Students with higher scores also attend higher quality colleges as measured by our college earnings index described above, as shown in Figure 1c. As with earnings, the correlations between test score percentiles and college outcomes are nearly linear throughout the distribution.

Figures 2a-2e present raw correlations between kindergarten test scores and other adult outcomes. Columns 4-8 in the first row of Table 4b show the corresponding regression coefficients. Conditional on student and parent demographic characteristics, a one percentile point increase in test scores increases the likelihood of owning a home before age 27 by 0.14 percentage points (Figure 2a), making a contribution to a 401(k) account by 0.10 percentage points (Figure 2b), and being married (Figure 2c) by 0.048 percentage points. Higher scoring students are also more mobile, as measured by living outside Tennessee prior to age 27 (Figure 2d), and live in higher SES neighborhoods in 2007 as measured by the percent of college graduates living in the ZIP code

[21] Similarly, Hanushek (2010) calculates that a 1 SD improvement in achievement would increase lifetime earnings by 13 percent.

(Figure 2e).

To analyze impacts on the other adult outcomes in a compact manner, we construct a summary index of the five outcomes shown in Figures 2a-2e. Following Kling, Liebman, and Katz (2007), we first standardize each outcome by subtracting its mean and dividing it by its standard deviation. We then sum the five standardized outcomes and divide by the standard deviation of the sum to obtain an index that has a standard deviation of 1. A higher value of the index represents more desirable outcomes. Students with higher KG test scores have higher outcomes as measured by the summary index, as shown in Figure 2f. Conditional on demographic characteristics, a one percentile increase in KG test scores is associated with an improvement of 0.49% of a standard deviation in the summary outcome index (Column 9 of Table 4b).

In Appendix Table 2, we document the same correlations for six subgroups: blacks and whites, males and females, and lower vs. higher income students (based on eligibility for free or reduced-price lunches). For all outcomes and all subgroups, the correlations remain large and significant, showing that KG test scores are strong predictors of adult outcomes very broadly. These strong correlations motivate the question of whether interventions that raise early childhood test scores also yield improvements in adult outcomes, which is the focus of the remainder of the paper.

IV Impacts of Observable Classroom Characteristics

In this section, we analyze the impacts of three features of classrooms that we can observe in our data – class size, teacher characteristics, and peer characteristics.

IV.A Class Size

We estimate the causal impacts of class size on adult outcomes using an intent-to-treat regression specification analogous to Krueger (1999):

$$y_{icnw} = \alpha_{nw} + \beta_1 \text{SMALL}_{cnw} + \beta_3 X_{icnw} + \varepsilon_{icnw} \quad (1)$$

where y_{icn} is an outcome such as earnings for student i randomly assigned to classroom c at school n in grade w. The variable $SMALL_{cnw}$ is an indicator for whether the student was assigned to a small class upon entry. Because the randomization occurred to classrooms within schools, we include school by entry wave fixed effects (α_{nw}) in all specifications. The vector X_{icnw} includes the student and parent demographic characteristics described above: a quartic in household income interacted with an indicator for whether the parents are ever married, 401(k)

savings, home ownership, mother's age at child's birth, and the student's gender, race, age (in days), and free lunch status (along with indicators for missing data). To examine the robustness of our results, we report the coefficient both with and without this vector of controls. The inclusion of these controls does not significantly affect the estimates, as expected given that the covariates are balanced across classrooms.[22]

As a reference, in column 1 of Table 5, we estimate (1) with the first observed test score (in the year of entry) as the outcome. Consistent with Krueger (1999), we find that students assigned to small classes score 4.8 percentile points higher on tests in the year they enter a participating school. Note that the average student assigned to a small class spent 2.27 years in a small class, while those assigned to a large class spent 0.13 years in a small class. On average, large classes had 22.6 students while small classes had 15.1 students. Hence, the impacts on adult outcomes below should be interpreted as effects of attending a class that is 33% smaller for 2.14 years.

College Attendance. We begin by analyzing the impacts of class size on college attendance. Figure 3a plots the fraction of students who attend college in each year from 1999 to 2007 by class size. In this and all subsequent figures, we adjust for school by entry wave effects to isolate the random variation of interest. To do so, we regress the outcome variable on school-by-wave dummies and the small class indicator in each tax year. We then construct the two series shown in the figure by requiring that the difference between the two lines equals the regression coefficient on the small class indicator in the corresponding year and the weighted average of the lines equals the sample average in that year.

Figure 3a shows that students assigned to a small class are more likely to attend college, particularly before age 25. As the cohort ages from 19 (in 1999) to 27 (in 2007), the attendance rate of both treatment and control students declines, consistent with patterns in the broader population. Because our measure of college attendance is based on tuition payments, it includes students who attend higher education institutions both part-time and full-time. Measures of college attendance around age 20 (two years after the expected date of high school graduation) are most likely to pick up full-time attendance to two-year and four-year colleges, while college attendance in later years may be more likely to reflect part-time enrollment. This could explain why the effect of class size becomes much smaller after age 25. We therefore analyze two measures of college attendance below: college attendance at age 20 and attendance at any point before age 27.

[22] Clustering the standard errors by classroom to account for potential class-level errors uniformly *reduces* the standard errors. We report the un-clustered standard errors to be conservative.

The regression estimates reported in Column 2 of Table 5 are consistent with the results in Figure 3a. Controlling for demographic characteristics, students assigned to a small class are 1.8 percentage points (6.7%) more likely to attend college in 2000. This effect is statistically significant with $p < 0.05$. Column 3 shows that students in small classes are 1.6 percentage points more likely to attend college at some point before age 27.

Next, we investigate how class size affects the quality of colleges that students attend. Using the earnings-based college quality measure described above, we plot the distribution of college quality attended in 2000 by small and large class assignment in Figure 3b. To adjust for school by wave effects, we compute residual college mean earnings from a regression on school by wave effects and plot the distribution of the residual within small and large classes, adding back the sample mean to facilitate interpretation of units. To show where the additional students who attend college in small classes go, the densities are scaled to integrate to the total college attendance rates for small and large classes. The excess density in the small class group lies primarily among the lower quality colleges, suggesting that the marginal students who were induced to attend college because of reduced class size enrolled in relatively low quality colleges. This is precisely what one would expect, as students who are on the margin of attending college are unlikely to end up attending top colleges.

Column 4 of Table 5 shows that students assigned to a small class attend colleges whose students have mean earnings that are $109 higher. That is, based on observed impacts on college attendance, we predict that students in small classes will be earning approximately $109 more per year at age 28. This earnings increase incorporates the extensive-margin of higher college attendance rates, because students who do not attend college are assigned the mean earnings of individuals who do not attend college in our index.[23] Conditional on attending college, students in small classes attend *lower* quality colleges on average because of the selection effect shown in Figure 3b.[24]

Earnings. Figure 3c shows the analog of Figure 3a for wage earnings. Earnings rise rapidly over time because many students are in college in the early years of the sample. Individuals in small classes have slightly higher earnings than those in large classes in most years. Column 5 of Table 5 shows that without controls, students who were assigned to small classes are estimated to earn $4 more per year on average between 2005 and 2007. With controls for demographic

[23] Alternative earnings imputation procedures for those who do not attend college yield similar results. For example, assigning these students the mean earnings of Tennessee residents or STAR participants who do not attend college generates larger estimates.

[24] Quantifying the effect of reduced class size on college quality for those who were already planning to attend college would require additional assumptions such as rank preservation.

characteristics, the point estimate of the earnings impact becomes -$124 (with a standard error of $325). Though the point estimate is negative, the upper bound of the 95% confidence interval is an earnings gain of $525 (3.3%) gain per year. As a result, the observed earnings impact is not statistically distinguishable from the $109 impact predicted by the college attendance impacts. If we were to predict the expected earnings gain from being assigned to a small class from the cross-sectional correlation between test scores and earnings reported in column 3 of Table 4a, we obtain an expected earnings effect of 4.8 percentiles $\times$ $94 = $451. The observed small class impact of -$124 is also statistically indistinguishable from this prediction.

In Appendix Table 3, we consider five alternative measures of wage earnings: (1) an indicator for having positive wage earnings in the years 2005-2007, (2) an indicator for having average wage earnings between 2005-2007 greater than the sample median ($12,553), (3) the within-sample percentile of a student's average wage earnings, (4) total household income, and (5) wage earnings in 2007, the least year in which we have wage earnings data. We find qualitatively similar impacts – point estimates close to zero with confidence intervals that include the predicted value from cross-sectional estimates – for all of these measures. The class size intervention, which raises test scores by 4.8 percentiles, is unfortunately not powerful enough to detect earnings increases of a plausible magnitude as of age 27. Because class size has significant impacts on college attendance, earnings effects might emerge in subsequent years, especially since college graduates have much steeper earnings profiles than non college graduates.

Other Outcomes. Column 6 of Table 5 shows that students assigned to small classes score 4.6 percent of a standard deviation higher in the summary outcome index defined above, an effect that is statistically significant with $p < 0.05$. This index combines information on savings behavior, home ownership, marriage rates, mobility rates, and residential neighborhood quality. In Appendix Table 4, we analyze the impacts of class size on each of the five outcomes separately. We find particularly large and significant impacts on the probability of having a 401(k), which can be thought of as a proxy for having a good job. This result is consistent with the view that students in small classes may have higher permanent income that could emerge in wage earnings measures later in their lifecycles. We also find positive effects on all the other components of the summary index, though these effects are not individually significant. As an additional robustness check, we analyze an alternative summary index that weights each of the five components by their impacts on wage earnings. We construct this index by regressing wage earnings on the five components in the cross-section and predicting wage earnings for each individual. We find significant impacts of

class size on this predicted-earnings summary index (column 6 of Appendix Table 4), confirming that our results are robust to the way in which the components of the summary index are weighted.

Overall, our reading of the evidence is that small class assignment in grades K-3 produces gains across a variety of outcomes in early adulthood. In Appendix Table 5, we document the heterogeneity of class size impacts across subgroups. We replicate the analysis of class size impacts in Table 5 for six groups: black and white students, males and females, and lower- and higher-income students (based on free lunch status). The point estimates of the impacts of class size are positive for most of the groups and outcomes, suggesting that the gains from small class size are widespread. The impacts on adult outcomes are somewhat larger for groups that exhibit larger test scores increases. For instance, black students assigned to small classes score 6.9 percentile points higher on their first observed test, are 5.3 percentage points more likely to ever attend college, and have an earnings increase of $250 (with a standard error of $494). However, virtually none of the differences in impacts across subgroups are statistically significant. Unfortunately, the STAR experiment is not powerful enough to detect heterogeneity in impacts of class size on adult outcomes.

IV.B Observable Teacher and Peer Effects

We estimate the impacts of observable characteristics of teachers and peers using specifications analogous to (1):

$$y_{icnw} = \alpha_{nw} + \beta_1 \text{SMALL}_{cnw} + \beta_2 z_{cnw} + \beta_3 X_{icnw} + \varepsilon_{icnw} \quad (2)$$

where z_{cnw} denotes a vector of teacher or peer characteristics for student i assigned to classroom c at school n in grade w. Because students were randomly assigned to classrooms, β_2 can be interpreted as the causal effect of the relevant teacher or peer characteristics on the outcome y. Note that we control for class size in these regressions, so that the variation identifying teacher and peer effects is orthogonal to that used above.

Teachers. We begin by examining the impact of teacher experience on scores and earnings. Figure 4a plots KG scores vs. the numbers of years of experience that the student's kindergarten teacher had at the time she taught his class. We adjust for school by entry wave effects by regressing the outcome and dependent variables on these fixed effects and computing residuals. The figure is a scatter plot of the residuals, with the sample means added back to facilitate interpretation of the axes. Figure 4a shows that students assigned to more experienced KG teachers have higher

test scores. The effect of experience on scores is roughly linear in the STAR experimental data, in contrast with other studies which find that the returns to experience drop sharply after the first few years. We return to this issue below.

Figure 4b replicates 4a for the earnings outcome. It shows that students who were randomly assigned to more experienced KG teachers have higher earnings at age 27. As with scores, the impact of experience on earnings in these data is roughly linear. Figure 4c characterizes the time path of the earnings impact. We divide teachers in two groups – those with experience above and below 10 years (since mean years of experience is 9.3 years). We then plot mean earnings for the students in the low- and high-experience groups by year, adjusting for school by wave effects as in Figure 3c. From 2000 to 2004 (when students are aged 20 to 24), there is little difference in earnings between the two curves. A gap opens starting in 2005; by 2007, students who had high-experience teachers in kindergarten are earning $1,104 more on average. As noted above, the emergence of the earnings impact in the mid 20s is to be expected because higher-achieving students are more likely to be in college in their early 20's, depressing their average level of earnings during that period.

Columns 1-3 of Table 6 quantify the impacts of teacher experience on scores and earnings, conditioning on the standard vector of student and parent demographic characteristics. Column 1 shows that students assigned to a teacher with than 10 years of experience score 3.2 percentile points higher on KG tests. Column 2 shows that these same students earn $1,093 more on average between ages 25 and 27 ($p < 0.05$). Column 3 estimates a linear model and shows that each extra year of teacher experience generates $57 more on average, consistent with Figure 4b.

The few other observable teacher characteristics in the STAR data (degrees, race, and progress on a career ladder) have no significant impact on scores or earnings. For instance, columns 1-3 of Table 6 show that the effect of teachers' degrees on scores and earnings is statistically insignificant. The finding that experience is the only observable measure that predicts teacher quality matches earlier studies of teacher effects (Hanushek 2010, Rockoff and Staiger 2010).

The causal impact of kindergarten teacher experience on earnings must be interpreted very carefully. Our results show that placing a child in a kindergarten class taught by a more experienced teacher yields improved outcomes. This finding does *not* imply that increasing a given teacher's experience will improve student outcomes. The difference in earnings of students with experienced teachers could be due to the intrinsic characteristics of experienced teachers rather than experience of teachers per se. For instance, teachers with more experience have selected to stay in the

profession and may be more passionate or more skilled at teaching. Alternatively, teachers from older cohorts may have been more skilled (Corcoran, Evans, and Schwab 2004, Hoxby and Leigh 2004, Bacolod 2007). The linear relationship between teacher experience and scores in the STAR data – unlike in other data sources – suggests that other factors correlated with experience may drive the observed impacts on scores and earnings. Therefore, we can conclude that early childhood teaching has a causal impact for long term outcomes, but cannot isolate the characteristics of teachers that drive the effect.

We restrict attention to kindergarten teachers in Figure 4 and Table 6 because teacher experience has little impact on test scores in grades 1-3 in the STAR data. Corresponding to this pattern, we also find no significant effect of teacher experience on earnings in grades 1-3. These findings are also consistent with the view that the experience of KG teachers may proxy for another unobserved characteristic of teachers that drives the observed impacts. For instance, selection patterns into kindergarten teaching across cohorts or over time might differ from those in later grades.

Peers. Better classmates could create an environment more conducive to learning, leading to improvements in adult outcomes. To test for such peer effects, we follow the standard approach in the recent literature by using linear-in-means specifications. We proxy for peer abilities (z) in (2) with the following exogenous peer characteristics: fraction black, fraction eligible for free or reduced-price lunch, fraction female, and mean age. Previous studies have shown that these characteristics are significant predictors of test scores. We replicate these findings in column 4 of Table 6, which regresses KG test scores on these four characteristics. Column 5 replicates column 4 with earnings as the dependent variable. The estimates on all four peer characteristics are very imprecise. For instance, the estimated effect of increasing the fraction of low-income peers by 1 percentage point is an earnings loss of $285, but with a standard error of $1593. In an attempt to obtain more power, we construct a single index of peer abilities by first regressing scores on the full set of parent and student demographic characteristics described above and then predicting peers' scores using this regression. However, as column 6 of Table 6 shows, even this single predicted peer score measure does not yield a precise estimate of peer effects on earnings; the confidence interval for a 1 percentile point improvement in peers' predicted test scores ranges from -$197 to $165.[25]

The STAR experiment lacks the power to measure the effects of observable peer characteristics on earnings precisely because the experimental design randomized students across classrooms. As a

[25]We find positive but insignificant impacts of teacher and peer characteristics on the other outcomes above, consistent with a general lack of power in observable characteristics (not reported).

result, it does not generate significant variation in mean peer abilities across classes. The standard deviation of mean predicted peer test scores (removing variation across schools and waves) is less than two percentile points.[26] This small degree of variation in peer abilities is adequate to identify effects on contemporaneous effects on test scores but proves to be insufficient to identify effects on outcomes twenty years later, which are presumably subject to much higher levels of idiosyncratic noise.

V Impacts of Unobservable Classroom Characteristics

Many unobserved aspects of teachers and peers could impact student achievement and adult outcomes. For instance, some teachers may generate greater enthusiasm among students or some peers might be particularly disruptive. To test whether such unobservable aspects of class quality have long-term impacts, we follow the analysis-of-variance approach taken in the modern literature on teacher effects. In particular, we test for "class effects" on scores and earnings by exploiting random assignment to classrooms. These class effects include the effects of teachers, peers, as well as any class-level shocks such as noise outside the room.

We begin by formalizing our approach to measuring unobserved class quality using a simple empirical model. We then characterize unobserved class effects using two techniques: an analysis of variance and a regression approach that uses peer test scores to measure unobserved class quality.

V.A A Statistical Model of Class Effects

Consider the following empirical model for test scores (s) at the end of the class and earnings or other adult outcomes (y):

$$s_{icn} = d_n + z_{cn} + a_{icn} \tag{3}$$

$$y_{icn} = \delta_n + \beta z_{cn} + \rho a_{icn} + \nu_{icn} \tag{4}$$

Here a_{icn} denotes intrinsic academic ability of student i in class c at school n. The error term ν_{icn} represents the component of intrinsic earnings ability that is uncorrelated with academic ability. The parameter ρ controls the correlation between intrinsic academic and earning ability. The school fixed effects d_n and δ_n capture school-level differences in achievement on tests and earnings

[26] Peer characteristics remain balanced even in grades 1-3 because there is no evidence of non-random sorting across classrooms or types after kindergarten. For instance, children of high income parents are no more likely to switch to small classes in later grades.

outcomes, e.g. due to variation in socioeconomic characteristics. Let I denote the number of students per class, C the number of classes per school, and N the number of schools.[27]

The variable z_{cn} denotes a class-level intervention, such as better teaching, smaller class size, or better peers. Such interventions affect both test scores and earnings. To simplify notation, we normalize the effect of z on test scores to 1. The effect of the intervention z on earnings is measured by the parameter β, which can be interpreted as the effect of an intervention that increases test scores by one percentile point on earnings. Different interventions affect scores and earnings in different ways, so β varies across interventions. Higher quality teachers may raise both earnings and test scores, leading to a positive β. At the other extreme, teaching purely to the test could increase scores without affecting earnings, leading to $\beta = 0$. Because of random assignment to classrooms, students' intrinsic abilities (a_{icn} and ν_{icn}) are orthogonal to z_{cn}.

For interventions z that are directly observable, one can estimate (3) and (4) directly using OLS, as we did in (1) and (2) to analyze the impacts of class size and observable teacher and peer attributes. We now describe two empirical methods of estimating the impacts of an intervention z_{cn} that we cannot observe, such as differences in unmeasured dimensions of teacher quality.

Strategy 1: Analysis of Variance. The simplest way to detect unobserved class-level variation in z_{cn} is to implement an analysis of variance (ANOVA). The ANOVA effectively decomposes the variation in y_{icn} into individual and class-level components, and tests for the significance of class-level variation. This test is equivalent to testing whether the outcome y varies across classes by more than what would be predicted by random variation in students across classrooms. Using a random class effects specification for (4), one can quantify the amount of variation in z_{cn} across classes within a school, $\sigma_c^2 = var(z_{cn} - z_n)$.

Although the ANOVA is useful for testing the null that there are no class effects, it has two limitations. First, it does not tell us whether the classes that raise student achievement on tests are those that have positive adult impacts. Given that the impacts of most educational interventions can only be measured by test scores in the short run, it is important to understand the covariance between impacts on scores and earnings. Second, in the STAR data, roughly half the students enter in grades 1-3 and are randomly assigned to classrooms at that point. Because only a small number of students enter each school in each of these later waves, we do not have the power to detect class effects in later grades. Therefore, we only have adequate precision to implement the ANOVA for

[27] We assume that I and C do not vary across classes and schools for presentational simplicity. Our empirical analysis accounts for variation in I and C across classrooms and schools, and the analytical results below are unaffected by such variation.

kindergarten classrooms. Motivated by these limitations, our second estimation strategy measures class quality using an index based on test scores that allows us to pool entrants in all waves.

Strategy 2: Peer-Score Measure of Class Quality. We develop an empirically observable proxy for class quality based on end-of-class test scores. Our approach is very similar to that of the modern teacher effects literature, which identifies teachers who repeatedly raise student achievement on tests as high quality teachers. Here, we exploit random assignment to classrooms and identify high quality classrooms as those that generate high average test scores. We present a heuristic sketch of our estimation strategy in the text; proofs are given in Appendix B. We first consider a case without peer effects by assuming $z_{cn} \perp a_{icn}$ for all i, and then discuss how peer effects affect our estimator.

Let s_{cn} denote the mean test score in class c and s_n denote the mean test score in school n. The mean test score in class c is

$$s_{cn} = \frac{1}{I}\sum_{j=1}^{I} s_{icn} = d_n + z_{cn} + \frac{1}{I}\sum_{j=1}^{I} a_{icn}$$

To simplify notation, assume that the mean value of z within a school is 0 ($z_n = 0$). Then the difference between mean test scores in a given class c and mean scores in the school is

$$\Delta s_{cn} = s_{cn} - s_n = z_{cn} + [\frac{1}{I}\sum_{j=1}^{I} a_{jcn} - \frac{1}{IC}\sum_{c=1}^{C}\sum_{j=1}^{I} a_{jcn}]. \tag{5}$$

Equation (5) shows that the difference in mean scores Δs_{cn} constitutes a (noisy) observable measure of class quality z_{cn}. The noise arises from variation in student abilities across classes. As the number of students grows large ($I \to \infty$), Δs_{cn} converges to the true underlying class quality z_{cn} given random assignment.

Equation (5) suggests that one could estimate the impact of class quality on earnings by substituting Δs_{cn} for z_{cn} in (4) and estimating a regression of the form:

$$y_{icn} = \alpha_n + b^M \Delta s_{cn} + \varepsilon_{icn}. \tag{6}$$

The OLS estimate $\widehat{b}^M$ is a consistent estimate of β as the number of students $I \to \infty$, but it is upward-biased with finite class size. In particular, a high ability student has both high earnings and raises mean class test scores Δs_{cn} in a class of finite size, generating an upward bias proportional to $\rho var(a_{icn})/I$. Therefore, even absent any true class effect ($\beta = 0$), $\mathbb{E}b > 0$ in (6).

One intuitive solution to eliminate the upward bias is to omit s_{icn} from the measure of class

quality for individual i (Angrist, Imbens, and Krueger 1999). We proxy for class quality using the leave-out mean (or jackknife) peer score measure

$$\Delta s_{cn}^{-i} = s_{cn}^{-i} - s_{n}^{-i}, \tag{7}$$

where

$$s_{cn}^{-i} = \frac{1}{I-1} \sum_{j=1, j\neq i}^{I} s_{jcn}$$

represents mean classmates' test scores and

$$s_{n}^{-i} = \frac{1}{IC-1} \sum_{k=1}^{C} \sum_{j=1, j\neq i}^{I} s_{jkn}$$

represents mean schoolmates' scores. The measure Δs_{cn}^{-i} answers the question: "How good are your classmates' scores compared with the classmates you could have had in your school?" Replacing Δs_{cn} by Δs_{cn}^{-i}, we estimate regressions of the following form:

$$y_{icn} = \alpha_n + b^{LM} \Delta s_{cn}^{-i} + \varepsilon_{icn}. \tag{8}$$

We show in Appendix B that using the leave-out mean measure of class quality Δs_{cn}^{-i} yields an OLS estimate whose expectation

$$\mathbb{E}\hat{b}^{LM} = \beta \times \frac{\frac{IC}{IC-1} var(z_{cn} - z_n)}{\frac{(IC)^2}{(IC-1)^2} var(z_{cn} - z_n) + var(a_{icn})(\frac{1}{I-1} - \frac{1}{I \cdot C-1})} \tag{9}$$

It follows that in equation (8), the coefficient on class quality is positive if and only if class quality has a causal impact on adult outcomes. Formally, $\mathbb{E}\hat{b}^{LM} > 0$ if and only if $\beta > 0$. As expected, the leave-out mean eliminates the upward bias in (6). However, b^{LM} is downward biased relative to β because Δs_{cn}^{-i} is a noisy measure of class quality with finite class size. In Appendix B, we calibrate the degree of this attenuation bias using the sample variance of test scores. We find that the empirical estimates of class quality effects reported below are attenuated relative to the true β by roughly 23%.[28]

We can include entrants in later waves when estimating (8) by defining Δs_{cn}^{-i} as the difference between mean end-of-year test scores for classmates and other peers in the student's grade in the year he entered a STAR school. With this definition, b^{LM} measures the extent to which class

[28] One could instead measure class quality using peers' adult outcomes directly; however, a test-score based measure is much less noisy because the variance of earnings across individuals is substantially greater than the variance of scores in practice ($var(\nu_{icn}) > var(a_{icn})$). Moreover, such a measure would not identify whether the classes that are high-quality as measured by impacts on scores are those that improve adult outcomes.

quality in the initial class of entry (to which students are randomly assigned) affects outcomes.[29]

Another method of overcoming the upward bias in (6) is to omit the own-observation using a split-sample approach (Angrist and Krueger 1995). This approach divides each class into two groups randomly and uses peer scores in the other group to proxy for class quality (see Appendix B). The split-sample estimator has properties similar to those of the leave-out mean estimator, but yields less precise and more attenuated estimates because it uses half the students to measure the quality of each class. We report estimates using the leave-out mean measure in the main text and, as a robustness check, replicate all of the results using the split-sample estimator in the appendix.

The result in (9) ignores variation in class quality due to peer effects because it assumes $z_{cn} \perp a_{icn}$. We extend the analysis to a standard linear-in-means model of peer effects by modelling class quality as

$$z_{cn} = t_{cn} + \frac{\theta}{I} \sum_{j} a_{jcn}$$

with $t_{cn} \perp a_{i'cn}$ for all i'. In this model, t_{cn} represents pure class effects independent of peer effects (e.g., a teacher effect) and the parameter $\theta \geq 0$ captures the strength of peer effects. When $\theta > 0$, a high ability student directly raises the scores and the earnings of his peers. Such peer effects bias b^{LM} upward because of the reflection problem (Manski 1993): a high ability student with high earnings also raises his peers scores, driving up the correlation between peer scores and own earnings. As a result, $\mathbb{E}\hat{b}^{LM} > \beta$ when $\theta > 0$. We cannot purge our simple leave-out mean estimator of this bias non-parametrically because it depends on the degree of peer effects θ. However, after reporting our baseline estimates, we bound the degree of reflection bias and find that the upper bound is an order of magnitude smaller than the class quality impacts we estimate. Abstractly, the reflection bias is of order $\frac{1}{I}$ and thus turns out to be relatively small in classes that have 20 students on average.

While we considered a model with a one-dimensional intervention z_{cn} above for expositional simplicity, our empirical proxy Δs_{cn}^{-i} should be interpreted as an omnibus measure of "class quality." The measure Δs_{cn}^{-i} incorporates both the effects of the teacher on students and variation in peer quality. It also incorporates any class-level shock that affects achievement, such as noise outside the room or disruptive interactions between particular students and teachers.

[29] There is a small downward bias in b^{LM} in later waves when Δs_{cn}^{-i} is defined based on comparisons with all peers in the school rather than peers who entered in the same entry wave. In practice, correcting for this bias by modifying (7) to include a comparison to peers in the same entry wave increases the coefficient estimates by less than 2%. We therefore opt to use the estimator in (8) for simplicity.

V.B Analysis of Variance

We implement the analysis of variance using regression specifications of the following form for students who enter the experiment in kindergarten:

$$y_{icn} = \alpha_n + \gamma_{cn} + \beta_3 X_{icn} + \varepsilon_{icn} \tag{10}$$

where y_{icn} is an outcome for student i who enters class c in school n in kindergarten and γ_{cn} is the class effect on the outcome.[30]

We first estimate (10) using a fixed-effects specification for the class effects γ_{cn}. Under the null hypothesis of no class effects, the class dummies should not be significant because of random assignment of students to classrooms. We test this null hypothesis using an F test for whether $\gamma_{cn} = 0$ for all c, n. This test is equivalent to testing whether the outcome y varies across classes by more than what would be predicted by random variation in students across classrooms. Note that we used exactly the same method in Section II.C to evaluate whether pre-determined variables such as the parental income or the student's race varied systematically across classrooms. Consistent with the randomized experimental design, the null hypothesis of no class effects was not rejected for any of the pre-determined variables, giving us reassurance in interpreting any class effects on outcomes as causal impacts of the classroom. To quantify the magnitude of the class effects, we compute the variance of γ_{cn} by estimating (10) using a random-effects specification. In particular, we assume that $\gamma_{cn} \tilde{} N(0, \sigma_c^2)$ and estimate the standard deviation of class effects σ_c.[31]

Table 7 reports p values from F tests and estimates of σ_c for test scores and earnings. Consistent with Nye, Konstantopoulos, and Hedges (2004) – who use an ANOVA to test for class effects on scores in the STAR data – we find highly significant class effects on KG test scores. Column 1 rejects the null hypothesis of no class effects on KG scores with $p < 0.001$. The estimated standard deviation of class effects on test scores is $\sigma_c = 8.77$, implying that a one standard deviation improvement in class quality raises student test scores by 8.77 percentiles (0.32 standard deviations). Note that this measure represents the impact of improving class quality by one SD of the *within-school* distribution because the regression specification includes school fixed effects.

Column 2 of Table 7 replicates the analysis in column 1 with 8th grade test scores as the outcome. We find no evidence that kindergarten classroom assignment has any lasting impact

[30] We omit γ_{cn} for one class in each school to avoid collinearity with the school effects α_n.

[31] Because we use a random-effects specification rather than the estimated fixed effects to calculate σ, there is no need to shrink the estimated class effects.

on achievement in 8th grade as measured by standardized test scores ($p = 0.42$). As a result, the estimated standard deviation of class effects on 8th grade scores is $\sigma_c = 0.00$. This evidence suggests that KG class effects fade out by grade 8, a finding that we revisit and explore in detail in Section VI.

Columns 3-6 of Table 7 implement the ANOVA for earnings (averaged from age 25-27). Column 3 implements the analysis without any controls besides school fixed effects. Column 4 introduces the full vector of parental and student demographic characteristics. Both specifications show statistically significant class effects on earnings ($p < 0.05$). The standard deviation of KG class effects on earnings in Column 4 (with controls) is $\sigma_c = \$1,520$. Assigning students to a classroom that is one standard deviation better than average for a single year in kindergarten generates an increase in earnings at age 27 of \$1,520 (9.6%) per year for each student. Although one cannot manipulate "class quality" directly, this substantial impact of class quality on adult earnings highlights the stakes at play in early childhood education.

Column 5 restricts the sample to students assigned to large classes, to test for class effects purely within large classrooms. This specification is of interest for two reasons. First, it isolates variation in class quality orthogonal to class size. Second, students in large classes were randomly reassigned to classrooms after kindergarten. Hence, column 5 specifically identifies the impact of kindergarten classrooms rather than a string of teachers and peers experienced over several years by a group of children who all started in the same KG class. Class quality continues to have a significant impact on earnings within large classes, showing that components of kindergarten class quality beyond size matter for earnings. Column 6 expands upon this approach by controlling for all observable classroom characteristics: indicators for small class, above-median teacher experience, black teacher, teacher with degree higher than a BA, and classmates' mean predicted score, constructed as in Column 6 of Table 6. The estimated σ_c falls by only \$66 relative to the specification in column 4, implying that most of the class effects are driven by features of the classroom that we cannot observe in our data.

V.C Effects of Class Quality on Scores and Earnings

To obtain more insight into why kindergarten classroom assignment affects adult outcomes, we use the peer-score measure of class quality defined above. In the baseline specifications, we include all students, regardless of the grade in which they entered the experiment. We analyze how the quality of the class to which students were randomly assigned upon entry into STAR affects

long-term outcomes. We estimate regression specifications of the following form:

$$y_{icnw} = \alpha_{nw} + \beta_1 \Delta s^{-i}_{cnw} + \beta_3 X_{icnw} + \varepsilon_{icnw} \tag{11}$$

Here y_{icnw} represents an outcome for student i who enters class c in school n in grade w. The key regressor of interest Δs^{-i}_{cnw} represents our leave-out mean measure of peer test scores for student i at the end of grade w, as defined in (7). After reporting the baseline results for the pooled sample, we test for differences in the impacts of class quality across grades K-3 by estimating (11) for separate waves. To maximize power, we first estimate (11) by pooling all potential sources of variation in class quality, including class size, teachers, and peers. We then control for the observable aspects of class quality analyzed above to isolate the impact of unobserved differences across classes.

We begin by characterizing the impact of class quality on test scores. Figure 5a plots each student's end-of-grade test scores vs. his initial class quality, as measured by his classmates' test scores minus his schoolmates' test scores. As in all other figures, we adjust for school by entry wave fixed effects. Figure 5a shows that children who are randomly assigned to higher quality classes upon entry – i.e. classes where their peers score higher on tests – have higher test scores at the end of the year. A one percentile increase in class quality is estimated to raise own test scores by 0.68 percentiles. Figure 5b replicates Figure 5a, changing the dependent variable to 8th grade test score. Consistent with the earlier ANOVA results, the gains from being in a higher quality kindergarten classroom fade out by grade 8. A one percentile increase in the quality of the class the student first entered in the STAR experiment raises 8th grade test scores by only 0.08 percentiles. Figure 6 uses the same design to evaluate the effects of class quality on adult wage earnings. Students assigned to a one percentile higher quality class have \$57.6 (0.4%) higher earnings on average from age 25-27.

We verify that our method of measuring class quality does not generate a mechanical correlation between peers scores and own outcomes using permutation tests. We randomly permute students across classrooms within schools and replicate (11). We use the t statistics on β_1 from the random permutations to form an empirical cdf of t statistics under the null hypothesis of no class effects. We find that fewer than 0.001% of the t statistics from the random permutations are smaller than the actual t-statistic on kindergarten test score in Figure 5a of 22.7. For the earnings outcome, fewer than 0.1% of the t statistics from the random permutations are smaller than the actual t-statistic of 3.55. These non-parametric permutation tests confirm that the p values obtained using parametric t-tests are accurate in our application.

Part of the relationship in Column 2 of Table 8a may be driven by reflection bias: high ability students raise their peers' scores and themselves have high earnings. This could generate a correlation between peer scores and own earnings even if class quality has no causal impact on earnings. However, the fact that end-of-kindergarten peer scores are not highly correlated with 8th grade test scores (Figure 5b) places a tight upper bound on the degree of this bias. Own ability has a strong impact on both KG and 8th grade scores: in the cross-section, a one percentile increase in KG test scores is associated with an 0.6 percentile increase in 8th grade score. In the presence of reflection bias, a high ability student in KG (who raises peer scores in KG) should also score highly on 8th grade tests, creating a spurious correlation between peer scores in KG and own 8th grade scores. Therefore, if end-of-KG peer scores have zero correlation with 8th grade scores, there cannot be any reflection bias. In Appendix B, we formalize this argument by deriving a bound on the degree of reflection bias in our linear-in-means model as a function of the empirical estimates in Table 8a and the cross-sectional correlations between test scores and earnings. If class quality has no causal impact on earnings ($\beta = 0$), the upper bound on the regression coefficient of earnings on class quality is \$9, less than 20% of our empirical estimate of \$57.6. Although this quantitative bound relies on the parametric assumptions of a linear-in-means model, it captures a more general intuition: the sharp "fade-out" of class quality effects on test scores rules out significant reflection bias in impacts of peer scores on later adult outcomes. Recall that the class quality estimates also suffer from a downward attenuation bias of 23%, the same magnitude as the upper bound on the reflection bias. We therefore proceed by using end-of-year peer scores as a simple proxy for class quality.

Figure 7a characterizes the time path of the impact of class quality on earnings, dividing classrooms in two groups of equal size – those with class quality above and below the median. The time pattern of the total class quality impact is similar to the impact of teacher experience shown in Figure 4c. Prior to 2004, there is little difference in earnings between the two curves, but a gap emerges starting in 2005. By 2007, students who were assigned to classes of above-median quality are earning \$930 (5.8%) more on average. Figure 7b shows the time path of the impacts on college attendance. Students in higher quality classes are more likely to be attending college in their early 20's, consistent with their higher earnings and steeper earnings trajectories in later years.

Table 8a quantifies the impacts of class quality on wage earnings using regressions with the standard vector of parent and student controls used above. Column 1 shows that conditional on the demographic characteristics, a one point percentile increase in class quality increases a student's

own test score by 0.66 percentile points. This effect is very precisely estimated, with a t-statistic of 27, because there is substantial variation in class quality as measured by peers' ex-post test scores. The standard deviation of our class quality measure is 9 percentiles. Because of the substantial increase in power, the peer-score based measure of class quality yields more precise estimates of the impact on adult outcomes than observable characteristics such as class size or teacher experience. Column 2 of Table 8a shows the effect of class quality on earnings. Conditional on demographic characteristics, a one percentile point increase in class quality increases earnings (averaged from 2005 to 2007) by \$50.61 per year, with a t-statistic of 3.3 ($p < 0.01$). To interpret the magnitude of this effect, note that a one standard deviation increase in class quality as measured by peer scores causes a \$455 (2.9%) increase in earnings at age 27. Appendix Table 3 replicates the specification in column 2 to show that class quality has positive impacts on all five alternative measures of wage earnings described above.

Column 3 of Table 8a isolates the variation in class quality that is orthogonal to class size by restricting the sample to students assigned to large classes. Class quality continues to have a significant impact on earnings within large classes, showing that components of class quality orthogonal to size matter for earnings. Column 4 expands upon this approach by controlling for all the observable classroom characteristics used in Column 6 of Table 7. The coefficient on end-of-year scores again does not change significantly, confirming that the impact of class quality on earnings is driven primarily by classroom characteristics that we do not observe in our data.

The preceding specifications pool grades K-3. Column 5 of Table 8a restricts the sample to kindergarten entrants and shows that a one percentile increase in KG class quality raises earnings by \$53. Because new entrants were randomly assigned to classes upon entry, we can estimate the effects of class quality in grades 1-3 using the subsample of students who entered STAR schools in grades 1-3. As above, we use peers' end-of-grade test scores to measure the quality of the class into which these new entrants were placed. The point estimate for new entrants shown in column 6 is similar to that in column 5, showing that class quality in grades 1-3 matters as much for earnings as class quality in kindergarten. These results suggest that the lessons of this study apply to early childhood classroom education rather than kindergarten per se.

The new entrants also allow us to estimate "value added" models of class quality, because we have measures of both end-of-year and prior year test scores for their peers. In column 7 of Table 8a, we replicate the specification in column 6 but include a control for class quality in the previous year, defined as the difference between mean classmates' and schoolmates' scores at the end of the

previous year. With this control, the coefficient on end-of-year test scores reported in column 7 can be interpreted as a measure of how *changes* in peer test scores in the current class affect adult outcomes. Conditioning on previous peer test scores reduces the coefficient on end-of-year peer scores modestly, implying that most of the impact of class quality is captured by changes in test scores over the current year.[32] This finding suggests that the variation in class quality is driven primarily by teachers or class-level shocks, such as interactions between teachers and peers ("classroom chemistry") in the current year.

Table 8b shows the impacts of class quality on other adult outcomes. These columns replicate the baseline specification for the full sample in column 2 of Table 8a. Columns 1 and 2 show that a 1 percentile improvement in class quality raises college attendance rates by 0.1 percentage points, both at age 20 and before age 27 ($p < 0.05$). Column 3 shows that a one percentile increase in class quality generates an \$9.3 increase in the college quality index ($p < 0.05$). Finally, column 4 shows that a one percentile point improvement in class quality leads to an improvement of 0.25% of a standard deviation in our outcome summary index ($p < 0.05$). Appendix Table 4b reports the impacts of class quality on each of the five outcomes separately and shows that the point estimates of the impacts are positive for all of the outcomes. In addition, class quality has a positive and significant impact on the predicted-earnings summary index defined in Section IV.A, showing that the results are robust to an alternative weighting of the five components of the summary index.

In Appendix Table 5b, we document the heterogeneity of class quality impacts across subgroups. The point estimates of the impacts of class size are positive for almost all the groups and outcomes, suggesting that the gains from class size are widespread. In Appendix Table 6, we proxy for class quality using the split-sample method, where classrooms are randomly divided into two groups and class quality is measured using mean peer scores in the other group. The results using the split-sample method are very similar to those obtained using the leave-out mean measure of class quality. In sum, there is robust evidence that improvements in early childhood class quality, as measured by peers' test scores, have substantial and lasting impacts on a broad range of adult outcomes.

[32]Lagged peer scores do not measure the quality of new entrants' previous class. Hence, the small coefficient on lagged peer scores does not imply that last year's class quality does not matter conditional on the current year's quality.

VI Fade-Out, Re-Emergence, and Non-Cognitive Skills

In this section, we explore why the impacts of class size and class quality in early childhood "fade out" on tests administered in later grades but re-emerge in adulthood. For simplicity, we focus on kindergarten entrants throughout this section and analyze the impacts of KG class quality on test scores and other outcomes in later grades.

We first document the fade-out effect using the class quality measure by estimating (11) for test scores in each grade with the standard vector of parent and student controls as well as school fixed effects. Figure 8a plots the estimated impacts of increasing KG class quality by one standard deviation on test scores in grades K-8. The standard deviation here refers to the distribution of class quality *within* schools because we include school fixed effects. A one SD increase in KG class quality increases end-of-kindergarten test scores by 6.27 percentiles, consistent with our findings above. In grade 1, students who were in a 1 SD better KG class score approximately 1.50 percentile points higher on end-of-year tests, an effect that is significant with $p < 0.001$. The effect gradually fades over time, and by grade 8 students who were in a better KG class no longer score significantly higher on tests. This fade-out effect is consistent with the rapid fade-out of teacher effects documented by Jacob, Lefgren, and Sims (2008), Kane and Staiger (2008), and others.

One potential explanation for the diminished test score impacts in later grades is that a given increase in test scores translates into a larger impact on earnings in later grades. If a one percentile increase in 8th grade test scores is more valuable than a one percentile increase in KG test scores, then the evidence in Figure 8a would not necessarily imply that the effects of early childhood education fade out. To evaluate this possibility, we convert the test score impacts to predicted earnings gains. We run separate OLS regressions of earnings on the test scores for each grade from K-8. We then multiply the class quality effect on scores shown in Figure 8a by the corresponding coefficient of earnings on scores from the OLS regression. Figure 8b plots the earnings impacts predicted by the test score gains in each grade that arise from attending a better KG class. The pattern in Figure 8b looks very similar to that in Figure 8a, showing that there is indeed substantial fade-out of the KG class quality effect in later grades. By 4th grade, one would predict less than a $50 per year gain in earnings from a better KG class based on observed test score impacts. These findings show that researchers who had examined only later test scores would indeed have been justified in concluding that early childhood education does not have long-lasting impacts.

The final point in Figure 8b shows the actual observed earnings impact of a one (within school)

standard deviation improvement in KG class quality. The actual impact of $483 is similar to what one would have predicted based on the improvement in KG test scores ($588). The impacts of early childhood education re-emerge in adulthood despite fading out on test scores in later grades. Because of this fade-out and re-emergence phenomenon, *contemporaneous* test scores – i.e. scores on tests taken at the end of the class year – are a good measure of the quality of the current class, but tests taken in subsequent years are not.

Non-Cognitive Skills. One potential explanation for fade-out and re-emergence is the acquisition of non-cognitive skills (Heckman 2000, Heckman and Rubinstein 2001, Heckman, Stixrud, and Urzua 2006, Cunha and Heckman 2008, Segal 2009). These studies use datasets such as the NELS and NLSY to show that non-cognitive skills (e.g., motivation or social skills) are highly correlated with adult outcomes even conditional on standardized test (e.g. math or reading) scores. We evaluate whether non-cognitive skills could explain our findings using data on non-cognitive measures collected for a subset of STAR students in grades 4 and 8. Previous studies have used these data to investigate whether class size affects non-cognitive skills (Finn et al. 1989, Dee and West 2008).[33] Here, we investigate the causal impacts of class quality (which includes class size as well as other factors) on non-cognitive outcomes and show how these impacts translate into earnings gains in adulthood.

Finn et al. (2007) and Dee and West (2008) describe the non-cognitive measures in the STAR data in detail; we provide a brief summary here. In grade 4, teachers in the STAR schools were asked to evaluate a random subset of their students on a scale of 1-5 on several behavioral measures, such as whether the student "annoys others." These responses were consolidated into four standardized scales measuring each student's effort, initiative, interest in the class, and disruptive behavior. In grade 8, Math and English teachers were asked to rate a subset of their students on a similar set of questions, which were again consolidated into four standardized scales. To obtain a measure analogous to our percentile measure of test scores, we construct percentile measures for these four scales and compute the average percentile score for each student. For 8th grade, we take the average of the math and English teacher ratings.

Among the 6,025 students who entered Project STAR in KG and whom we match in the IRS data, we have data on non-cognitive skills for 1,671 (28%) in grade 4 and 1,780 (30%) in grade

[33]These studies find mixed evidence on the impact of class size on non-cognitive skills: statistically significant impacts are detected in grade 4, but not in grade 8. We find similar, imprecisely estimated results on class size (consistent with our imprecise estimates of earnings impacts), but obtain more precise estimates using our more powerful class quality measure.

8. The availability of non-cognitive measures for only a subset of the students who could be tracked until grade 8 naturally raises concerns about selective attrition. Dee and West (2008) investigate this issue in detail, and we replicate their findings with our expanded set of parental demographic characteristics. We find no significant differences in the probability of having non-cognitive data in grade 8 across students in different kindergarten classrooms or class types (small vs. large). Replicating the tests of randomization in Table 2 for the subsample who have non-cognitive measures in grade 8 also reveals no significant differences in pre-determined student and parent characteristics across classrooms or types. In grade 4, non-cognitive data are significantly more likely to be available for students who were assigned to small classes than large classes. However, there are no significant differences in pre-determined student and parent characteristics across classrooms or types within the subsample who have non-cognitive data in grade 4. This suggests that attrition from the sample was random at least in terms of observed characteristics. This evidence supports the view that differences in observed non-cognitive outcomes across classrooms are due to the causal impact of class quality.

We begin our analysis of non-cognitive skills by estimating the cross-sectional correlation between non-cognitive outcomes and earnings. Column 1 of Table 9 shows that a 1 percentile improvement in non-cognitive measures in grade 4 is associated with a $106 gain in earnings conditional on the standard vector of parent and student demographic characteristics used above and school by wave fixed effects. Column 2 shows that controlling for math and reading test scores in grade 4 reduces the estimated impact of non-cognitive scores only slightly, to $88 per percentile. This shows that non-cognitive measures have considerable predictive power for adult outcomes beyond what is measured on standardized tests. In contrast, column 3 shows that non-cognitive skills in grade 4 are relatively weak predictors of 8th grade test scores when compared with math and reading scores in 4th grade. Because non-cognitive skills appear to be correlated with earnings through channels that are not picked up by subsequent standardized tests, they could explain fade-out and re-emergence.

To test this mechanism, we investigate the causal impacts of KG class quality on non-cognitive skills in grade 4 and 8. As a reference, Column 4 of Table 9 shows that a 1 percentile improvement in KG class quality increases a student's test scores in grade 4 by a statistically insignificant 0.05 percentiles. In contrast, column 5 shows that the same improvement in KG class quality generates a statistically significant increase of 0.15 percentiles in the index of non-cognitive measures in

grade 4. Columns 6 and 7 replicate columns 3 and 4 for grade 8.[34] Again, KG class quality does not have a significant impact on 8th grade test scores but has a significant impact on non-cognitive measures. Finally, columns 8 and 9 show that the experience of the student's teacher in kindergarten – which we showed above also impacts earnings – has an insignificant impact on test scores but a significant impact on non-cognitive measures in 8th grade. Appendix Table 7 breaks down the constituent components of the aggregate non-cognitive score. All four non-cognitive measures are highly correlated with earnings. Students in higher quality classes exhibit persistent improvements on all four components, suggesting that higher quality KG classes lead to broad improvements in non-cognitive skills.

We can translate the impacts on non-cognitive skills into predicted impacts on earnings following the method in Figure 8b. We regress earnings on the non-cognitive measure in grade 4, conditioning on demographic characteristics, and obtain an OLS coefficient of $101 per percentile. Multiplying this OLS coefficient by the estimated impact of class quality on non-cognitive skills in grade 4, we predict that a 1 SD improvement in KG class quality will increase earnings by $139. The same exercise for 4th grade test scores yields a predicted earnings gain of $40. Similar calculations for 8th grade imply a predicted earnings gain of $166 through improvements in non-cognitive skills compared with $86 through skills measured by standardized tests. In both cases, improvements in non-cognitive skills explain a larger share of actual earnings gains ($481) than improvements in cognitive performance in later grades.

The impacts of non-cognitive skills on later scores are much smaller than on earnings. Following the methodology above, a one standard deviation increase in class quality is predicted to raise 8th grade test scores by 0.47 percentiles based on its observed impacts on non-cognitive skills in grade 4 and the cross-sectional correlation between grade 4 non-cognitive skills and grade 8 test scores. This predicted impact is quite close to the actual impact of class quality on 8th grade scores of 0.57 percentiles. Hence, the impacts of class quality on non-cognitive skills is consistent with both fade-out on scores and re-emergence on adult outcomes. Intuitively, a better kindergarten classroom might simultaneously increase performance on end-of-year tests and improve *untested* non-cognitive skills. For instance, a KG teacher who is able to make her students memorize vocabulary words may instill social skills in the process of managing her classroom successfully. These non-cognitive skills may not be well measured by standardized tests, leading to very rapid fadeout immediately

[34] Because non-cognitive measures were collected for a random subset of students, we use the full sample in columns 4 and 6 to increase precision. The point estimates on test score impacts are similar for the subsample of students for whom non-cognitive data are available.

after KG as in Figure 8a. However, these skills could still have returns in the labor market.

Although non-cognitive skills provide a plausible explanation of the data, we caution that the evidence cannot be viewed as definitive proof of the importance of non-cognitive skills for three reasons. First, as noted above, attrition is a much more serious concern in the analysis of non-cognitive outcomes. While attrition appears to be random based on observables, there may be unobserved differences in attriters across classrooms. Second, though we find significant impacts of KG class quality on non-cognitive skills in grades 4 and 8, the effects are imprecisely measured. In particular, one cannot reject the hypothesis that the percentile impacts on test scores and the non-cognitive measures are the same. Third, and most importantly, our analysis does not show that manipulating non-cognitive skills directly has causal impacts on adult outcomes. We have shown that high quality KG classes improve both non-cognitive skills and adult outcomes, but the mechanism through which adult outcomes are improved could run through another channel that is correlated with the acquisition of non-cognitive skills. It would be very valuable to analyze interventions that target non-cognitive skills directly in future work.

VII Conclusion

The impacts of education have traditionally been measured by achievement on standardized tests. This paper has shown that many of the interventions that raise test scores – such as reduced class size or better teachers – also improve long-term outcomes. Students who were randomly assigned to higher quality classrooms in grades K-3 earn more, are more likely to attend college, save more for retirement, and live in better neighborhoods. While the quality of education is best judged by directly measuring its impacts on such outcomes, our analysis suggests that *contemporaneous* (end-of-year) test scores are a reasonably good measure of the quality of a classroom. Students who were in better classrooms in grades K-3 do not do much better on standardized tests in later grades. Improvements in non-cognitive skills may explain why the impacts of early childhood education fade-out in later grades and then re-emerge in adulthood.

We conclude by using our empirical estimates to calculate the benefits of various policy interventions. The calculations that follow should be interpreted as rough approximations that convey the order-of-magnitude of the impacts because they rely on the following strong assumptions. First, following Krueger (1999), we assume a 3% annual discount rate and discount all earnings streams back to age 6, the point of the intervention. Second, we use the mean wage earnings of a random sample of the U.S. population in 2007 as a baseline earnings profile over the lifecycle. Third, be-

cause we can only observe earnings impacts up to age 27, we must make an assumption about the impacts after that point. We assume that the percentage gain observed at age 27 remains constant over the lifecycle. This assumption may understate the total benefits because the earnings impacts appear to grow over time (Figures 4c, 7a), as college graduates have steeper earnings profiles. Finally, our calculations ignore non-monetary returns to education such as reduced crime. They also ignore general equilibrium effects: increasing the education of the population at large would increase the supply of skilled labor and may depress wage rates for more educated individuals, reducing total social benefits. Under these assumptions, we calculate the present-value earnings gains for a classroom of 20 students from three interventions: improvements in classroom quality, reductions in class size, and improvements in teacher quality.

(1) Class Quality. The random-effects estimate reported in column 4 of Table 7 implies that increasing class quality by one standard deviation of the distribution *within* schools raises earnings by $1,520 (9.6%) at age 27.[35] Under the preceding assumptions, this translates into a lifetime earnings gain of approximately $39,100 for the average individual. This implies a present-value benefit of $782,000 for improving class quality by one within-school standard deviation for a *single* year. This $782,000 figure includes all potential benefits from an improved classroom environment, including better peers, teachers, and random shocks. Importantly, part of these pre-tax earnings gains may not accrue to the individual due to increased tax liabilities. Because the STAR sample has relatively low incomes at age 27, we find no significant impacts of the interventions on tax liabilities, as shown in Table 10. However, our results suggest that in the longer run, improvements in early childhood education could potentially have important fiscal returns for the government.

The class quality calculation is useful primarily for understanding the stakes at play in early childhood education. It is less helpful from a policy perspective because one cannot implement interventions that directly improve classroom quality. This motivates the analysis of class size and better teachers, two factors that contribute to classroom quality.

(2) Class Size. We calculate the benefits of reducing class size by 33% in two ways. The first method uses the estimated earnings gain from being assigned to a small class reported in column 5 of Table 5. The point estimate of $4 in Table 5 translates into a lifetime earnings gain from reducing class size by 33% for one year of $103 in present value per student, or $2,057 for a class that originally had twenty students. But this estimate is imprecise: the 95% confidence interval

[35]The standard deviation of the distribution of class quality including variation across schools is presumably significantly larger.

for the lifetime earnings gain of reducing class size by 33% for one year ranges from -$17,500 to $17,700 per child. Moreover, the results for other measures such as college attendance suggest that the earnings and tax revenue impact may be larger in the long run.

To obtain more precise estimates, we predict the gains from class size reduction using the estimated impact of classroom quality on scores and earnings. We estimate that a 1 percentile increase in class quality raises test scores by 0.66 percentiles and earnings by $50.6. This implies an earnings gain of $76.67 per percentile (or 13.1% per standard deviation) increase in test scores. We make the strong assumption that the ratio of earnings gains to test score gains is the same for changes in class size as it is for improvements in class quality more generally.[36] Under this assumption, smaller classes (which raised test scores by 4.8 percentiles) are predicted to raise earnings by $4.8 \times \$76.7 = \$368$ (2.3%) at age 27. This calculation implies a present value earnings gain from class size reduction of $9,460 per student and $189,000 for the classroom.

Calculations analogous to those in Krueger (1999) imply that the average cost per child of reducing class size by 33% for 2.14 years (the mean treatment duration for STAR students) is $9,355 in 2009 dollars.[37] Our second calculation suggests that the benefit of reducing class size might outweigh the costs. However, we must wait for more time to elapse before we can determine whether the predicted earnings gains based on the class quality estimates are in fact realized by those who attended smaller classes.

(3) Teachers. We calculate the benefits of improving teacher quality in two ways. The first method uses the estimated earnings gain of $57 from being assigned to a teacher with one year of extra experience. The standard deviation of teacher experience in our sample is 5.8 years. Hence, a one standard deviation increase in teacher experience raises earnings by $331 (2.1%) at age 27. This translates into a lifetime earnings gain of $8,500 in present value, or $170,000 for a class of twenty students.

The limitation of the preceding calculation is that it is based upon only one observable aspect of teacher quality. To incorporate other aspects of teacher quality, we again develop a prediction based on the impacts of class quality on scores and earnings. Rockoff (2004), Rivkin, Hanushek, and Kain (2005), and Kane and Staiger (2008) use datasets with repeated teacher observations to estimate

[36]This assumption clearly does not hold for all types of interventions. As an extreme example, raising test scores by cheating would be unlikely to yield an earnings gain of $77 per percentile improvement in test scores. The $77 per percentile measure should be viewed as a prior estimate of the expected gain when evaluating interventions such as class size or teacher quality for which precise estimates of earnings impacts are not yet available.

[37]This cost is obtained as follows. The annual cost of school for a child is $8,848 per year. Small classes had 15.1 students on average, while large classes had 22.56 students on average. The average small class treatment lasted 2.14 years. Hence, the cost per student of reducing class size is (22.56/15.1-1)*2.14*8848= $9,355.

that a one standard deviation increase in teacher quality raises test scores by approximately 0.2 standard deviations (5.4 percentiles).[38] Under the strong assumption that the ratio of earnings gains to test score gains is the same for changes in teacher quality and class quality more broadly, this translates into an earnings gain of $5.4 \times \$76.7 = \$416$ (2.6%) at age 27. This implies a present-value earnings gain of $10,700 per student. A one standard deviation improvement in teacher quality in a single year generates earnings gains of $214,000 for a class of twenty students.

The magnitude of the estimated impacts of teachers on earnings suggests that good teachers can create great social value and hence increased long-term tax revenue, perhaps several times larger than current teacher salaries.[39] However, our results do not have direct implications for optimal teacher salaries or merit pay policies. An analogy with executive compensation might be helpful in understanding this point. CEOs' decisions have large impacts on the firms they run, and hence can create or destroy large amounts of economic value. But this does not necessarily imply that increasing CEO compensation or pay-for-performance would improve CEO decisions. Analogously, our analysis shows that good teachers may create tremendous value in terms of increased earnings, but does not tell us whether higher salaries or merit pay would improve teacher quality.

Relative to efforts that seek to improve the quality of teachers, class size reductions have the important advantage of being much more well-defined and straightforward to implement. However, our findings on the importance of teacher quality caution that reductions in class size must be implemented carefully to generate improvements in outcomes. If schools are forced to reduce teacher and class quality along other dimensions when reducing class size, the net gains from class size reduction may be diminished.

This paper is a first step that documents the long-term impacts of early childhood classroom environments using simple quasi-experimental methods. An important direction for future work is to directly identify the relative contributions of teachers, peers, and other aspects of classrooms to long-term outcomes. One promising approach is to exploit variation in class size to separate peer and teacher effects, following the innovative method developed by Graham (2008).[40] Another strategy would be to estimate teacher effects on earnings using datasets that cover multiple

[38] We use estimates of the impacts of teacher quality on scores from other studies to predict earnings gains because we do not have repeat observations on teachers in our data. In future work, it would be extremely valuable to link datasets with repeat observations on teachers to administrative data on students in order to measure teachers' impacts on earnings directly.

[39] According to calculations from the 2006-2008 American Community Survey, the mean salary for elementary and middle school teachers in the U.S. was $39,164 (in 2009 dollars).

[40] This strategy can be implemented within the STAR data itself by specifying a parametric model of peer effects. We defer such an analysis to future work in the interest of space.

classrooms per teacher, such as the Tennessee database used by Sanders and Horn (1998).

While further research is needed to determine the best ways to improve classroom quality, the broad message of these calculations is that children who attend higher quality schools fare substantially better as adults. In the United States, the current property-tax system of school finance gives higher income families access to better public schools on average. This system could amplify inequality, as disadvantaged children generally attend lower quality, resource-constrained schools. Our analysis of the long-term impacts of Project STAR suggests that improving early childhood education in disadvantaged areas may significantly reduce poverty and inequality in the long run.

Our study also generates valuable benefits for tax policy and tax administration. First, it shows that tax data can be uniquely valuable for analyzing the effects of a policy intervention (such as early childhood education) on long-term outcomes such as earnings, home ownership, or savings, that are critically important for future tax revenue. The type of cost and benefit analysis we have outlined in conclusion is a critical input in the evaluation of actual and prospective government policies. Second, our study demonstrates that the administration of tax credits for higher education not only provide monetary support for education but also generates data that provide a unique and comprehensive snapshot of college enrollment in the United States. Our study shows that those data can be used for research and are ideally suited to evaluate the effects and the administration of many government policies.

Appendix A: Algorithm for Matching STAR Records to Tax Data

STAR records were matched to tax data using social security number (SSN), date of birth, gender, name, and STAR elementary school ZIP code. Note that STAR records do not contain all the same information. Almost every STAR record contains date of birth, gender, and last name. Some records contain no SSN while others contain multiple possible SSNs. Some records contain no first name. A missing field yielded a non-match unless otherwise specified.

We first discuss the general logic of the match algorithm and then document the routines in detail. The match algorithm was designed to match as many records as possible using variables that are *not* contingent on ex post outcomes. SSN, date of birth, gender, and last name in the tax data are populated by the Social Security Administration using information that is not contingent on ex post outcomes. First name and ZIP code in tax data are contingent on observing some ex post outcome. First name data derive from information returns, which are typically generated after an adult outcome like employment (W-2 forms), college attendance (1098-T forms), and mortgage interest payment (1098 forms). The ZIP code on the claiming parent's 1040 return is typically from 1996 and is thus contingent on the ex post outcome of the STAR subject not having moved far from her elementary school by age 16.

89.8% of STAR records were matched using only ex ante information. The algorithm first matched as many records as possible using only SSN, date of birth, gender, and last name. It then used first name only to *exclude* candidate matches based on date of birth, gender, and last name, often leaving only one candidate record remaining. Because that exclusion did not condition on an information return having been filed on behalf of that remaining candidate, these matches also did not condition on ex post outcomes.

The match algorithm proceeded as follows, generating seven match types denoted A through G. The matches generated purely through ex-ante information are denoted A through E below and account for 89.8% of STAR records. Matches based on ex-post-information are denoted F and G below and constitute an additional 5.4% of STAR records. The paper reports results using the full 95.0% matched sample, but all the qualitative results hold in the 89.8% sample matched using only ex ante information.

1. Match STAR records to tax records by SSN. For STAR records with multiple possible SSNs, match on all of these SSNs to obtain a set of candidate tax record matches for each STAR record with SSN information. Each candidate tax record contains date of birth, gender, and first four letters of every last name ever assigned to the SSN.

    - Match Type A. Keep unique matches after matching on first four letters of last name, date of birth, and gender.
    - Match Type B. Refine non-unique matches by matching on either first four letters of last name or on "fuzzy" date of birth. Then keep unique matches. Fuzzy date of birth requires the absolute value of the difference between STAR record and tax record dates of birth to be in the set {0,1,2,3,4,5,9,10,18,27} in days, in the set {1,2} in months, or in the set {1} in years. This set was chosen to reflect common mistakes in recorded dates of birth, such as being off by one day (e.g. 12 vs. 13) or inversion of digits (e.g. 12 vs. 21).

2. Match residual unmatched STAR records to tax records by first four letters of last name, date of birth, and gender.

    - Match Type C. Keep unique matches.

- Match Type D. Refine non-unique matches by excluding candidates who have a first name issued on information returns (e.g. W-2 forms, 1098-T forms, and various 1099 forms) that does not match the STAR first name on first four letters when the STAR first name is available. Then keep unique matches.
- Match Type E. Refine residual non-unique matches by excluding candidates who have SSNs that, based on SSN area number, were issued from outside the STAR region (Tennessee and neighboring environs). Then keep unique matches.
- Match Type F. Refine residual non-unique matches by keeping unique matches after each of the following additional criteria is applied: require a first name match when STAR first name is available, require the candidate tax record's SSN to have been issued from the STAR region, and require the first three digits of the STAR elementary school ZIP code to match the first three digits of the ZIP code on the earliest 1040 return on which the candidate tax record was claimed as a dependent.

3. Match residual unmatched STAR records to tax records by first four letters of last name and fuzzy date of birth.

   - Match Type G. Keep unique matches after each of several criteria is sequentially applied. These criteria include matches on first name, last name, and middle initial using the candidate tax record's information returns; on STAR region using the candidate tax record's SSN area number; and between STAR elementary school ZIP code and ZIP code on the earliest 1040 return on which the candidate tax record was claimed as a dependent.

The seven match types cumulatively yielded a 95.0% match rate:

Match type	Frequency	Percent	Cumulative percent
A	7036	60.8%	60.8%
B	271	2.3%	63.1%
C	699	6.0%	69.2%
D	1391	12.0%	81.2%
E	992	8.6%	89.8%
F	299	2.6%	92.4%
G	304	2.6%	95.0%

Identifiers such as names and SSN's were used solely for the matching procedure. After the match was completed, the data were de-identified (i.e., individual identifiers such as names and SSNs were stripped) and the statistical analysis was conducted using the de-identified dataset.

Appendix B: Derivations for Measurement of Unobserved Class Quality

This appendix derives the estimators discussed in the empirical model in Section V and quantifies the degree of attenuation and reflection bias. We first use equations (3) and (4) to define average of test scores and earnings within each class c and school n:

$$\begin{aligned} s_{cn} &= d_n + z_{cn} + a_{cn} \\ y_{cn} &= \delta_n + \beta z_{cn} + \rho a_{cn} + \nu_{cn} \\ s_n &= d_n + z_n + a_n \\ y_n &= \delta_n + \beta z_n + \rho a_n + \nu_n. \end{aligned}$$

We can then define variables demeaned within schools as

$$\begin{aligned} s_{icn} - s_n &= z_{cn} - z_n + a_{icn} - a_n \\ \Delta s_{cn} \equiv s_{cn} - s_n &= z_{cn} - z_n + a_{cn} - a_n, \\ y_{icn} - y_n &= \beta(z_{cn} - z_n) + \rho(a_{icn} - a_n) + \nu_{icn} - \nu_n \\ y_{cn} - y_n &= \beta(z_{cn} - z_n) + \rho(a_{cn} - a_n) + \nu_{cn} - \nu_n. \end{aligned}$$

Recall that a_{icn} and ν_{icn} are each iid and independent of each other and z_{cn}. Let $\sigma^2 = var(a_{icn})$. We assume in parts 1-3 below that $z_{cn} \perp a_{icn}$, i.e. there are no peer effects.

1. Mean score estimator. The simplest proxy for class quality is the average test score within a class. Since we include school fixed effects in all specifications, s_{cn} is equivalent to Δs_{cn} as defined above. Therefore, consider the following (school) fixed effects OLS regression:

$$y_{icn} = \alpha_n + b^M \Delta s_{cn} + \varepsilon_{icn}. \tag{12}$$

The coefficient estimate has expectation

$$\mathbb{E}\hat{b}^M = \frac{cov(y_{icn} - y_n, s_{cn} - s_n)}{var(s_{cn} - s_n)},$$

which we can rewrite as

$$\begin{aligned} \mathbb{E}\hat{b}^M &= \frac{cov\left(\beta(z_{cn} - z_n) + \rho(a_{icn} - \frac{\sum_k \sum_j a_{jkn}}{I \cdot C}), z_{cn} - z_n + \frac{\sum_j a_{jcn}}{I} - \frac{\sum_k \sum_j a_{jkn}}{I \cdot C}\right)}{var\left(z_{cn} - z_n + \frac{\sum_j a_{jcn}}{I} - \frac{\sum_k \sum_j a_{jkn}}{I \cdot C}\right)} \\ &= \frac{\beta var(z_{cn} - z_n) + \rho\sigma^2 \frac{C-1}{IC}}{var(z_{cn} - z_n) + \sigma^2 \frac{C-1}{IC}}. \end{aligned}$$

Even absent class effects ($\beta = 0$), we have $\mathbb{E}\hat{b}^M > 0$ if I is finite and $\rho > 0$. With finite class size, b^M is upward-biased due to the correlation between wages and own-score, which is included within the class quality measure.

2. Leave-out mean estimator. The second estimator we consider — and the one we primarily use in the empirical analysis — is a leave-out mean. Consider the OLS regression with school fixed effects

$$y_{icn} = \alpha_n + b^{LM} \Delta s_{cn}^{-i} + \varepsilon_{icn}. \tag{13}$$

where $\Delta s_{cn}^{-i} = s_{cn}^{-i} - s_n^{-i}$ is defined as in (7). The coefficient b^{LM} has expectation

$$\mathbb{E}\hat{b}^{LM} = \frac{cov(y_{icn} - y_n, s_{cn}^{-i} - s_n^{-i})}{var(s_{cn}^{-i} - s_n^{-i})},$$

which we can rewrite as

$$\begin{aligned}\mathbb{E}\hat{b}^{LM} &= \frac{cov\left(\beta(z_{cn}-z_n)+\rho(a_{icn}-a_n), \frac{IC}{IC-1}(z_{cn}-z_n)+\frac{1}{I-1}\sum_{j\neq i}a_{jcn}-\frac{1}{IC-1}\sum_k\sum_{j\neq i}a_{jkn}\right)}{var\left(\frac{IC}{IC-1}(z_{cn}-z_n)+\frac{1}{I-1}\sum_{j\neq i}a_{jcn}-\frac{1}{IC-1}\sum_k\sum_{j\neq i}a_{jkn}\right)} \\ &= \beta\times\frac{\frac{IC}{IC-1}var(z_{cn}-z_n)}{\frac{(IC)^2}{(IC-1)^2}var(z_{cn}-z_n)+\frac{\sigma^2}{I-1}-\frac{\sigma^2}{I\cdot C-1}}\end{aligned}$$

Hence, $\mathbb{E}\hat{b}^{LM}=0$ if and only if $\beta var(z_{cn}-z_n)=0$ (no class effects) even when I and C are finite.[41] However, b^{LM} is attenuated relative to β because peer scores are a noisy measure of class quality.

Quantifying the degree of attenuation bias. We can quantify the degree of attenuation bias by using the within-class variance of test scores as an estimate of σ^2. First, note that:

$$\begin{aligned}\widehat{var}(z_{cn}-z_n) &= \frac{(IC-1)^2}{(IC)^2}\left[\widehat{var}\left(s_{cn}^{-i}-s_n^{-i}\right)-\left(\frac{\hat{\sigma}^2}{I-1}-\frac{\hat{\sigma}^2}{I\cdot C-1}\right)\right] \\ &= \frac{(83.63)^2}{(84.73)^2}\left[81.75-\left(\frac{437.4}{19.07}-\frac{437.4}{83.63}\right)\right] \\ &= 62.39\end{aligned}$$

where we use the sample harmonic means for IC, $IC-1$, and $I-1$ because the number of students in each class and school varies across the sample. This implies an estimate of bias of

$$\frac{\frac{83.63}{84.73}62.39}{\frac{(83.63)^2}{(84.73)^2}62.39+\frac{437.4}{19.07}-\frac{437.4}{83.63}}=0.773.$$

That is, b^{LM} is attenuated relative to β by 22.7%.

3. Split sample estimator. To construct the split-sample proxy for class quality used in Appendix Table 6, we randomly split each class into two groups, $g=1,2$. We then use the average end-of-year test score in one group as the class quality measure for students in the *other* group. Formally, define mean peer scores in the other group as

$$s_{cn}^{-g}=d_n+z_{cn}+\frac{2}{I}\sum_{j\in -g}a_{jcn}$$

where g denotes the students group in class c and school n. By construction, a student's class quality does not contain her own test score. Now consider the following OLS regression that includes school-by-group fixed effects:

$$y_{icn}=\alpha_{ng}+b^{SS}s_{cn}^{-g}+\varepsilon_{icn}. \tag{14}$$

[41] Equation (13) is similar to specifications used to estimate peer effects, but uses ex-post peer outcomes on the right hand side. Regression specifications used to estimate peer effects typically control for the individual's characteristic (e.g. age) and examine the effect of mean ex-ante peer characteristics (e.g. mean age) on outcomes. We do not control for own scores because end-of-year scores are endogenous: the effect of peer scores (which measures class quality) on outcomes runs through impacts on own scores. Without controlling for own scores, there is a downward bias in proxying for class quality purely using classmates' scores (s_{cn}^{-i}), because above-average students mechanically have below-average classmates. The difference $s_{cn}^{-i}-s_n^{-i}$ eliminates this mechanical bias.

The coefficient b^{SS} has expectation

$$\mathbb{E}\hat{b}^{SS} = \frac{cov(y_{icn} - y_n, s_{cn}^{-g} - s_{ng}^{-g})}{var(s_{cn}^{-g} - s_{ng}^{-g})}$$

where the ng fixed effects average over all classes in n and students in the g group, so that

$$s_{ng}^{-g} = d_n + z_n + \frac{2}{IC}\sum_{c'}\sum_{i' \in -g} a_{i'c'n}.$$

It follows that

$$\begin{aligned}
\mathbb{E}\hat{b}^{SS} &= \frac{cov\left(\beta(z_{cn} - z_n) + \rho(a_{icn} - a_n), z_{cn} - z_n + \frac{2}{I}\sum_{j\in -g} a_{jcn} - \frac{2}{IC}\sum_{j\in -g,k} a_{jkn}\right)}{cov\left(z_{cn} - z_n + \frac{2}{I}\sum_{j\in -g} a_{jcn} - \frac{2}{IC}\sum_{j\in -g,k} a_{jkn}\right)} \\
&= \beta \times \frac{var(z_{cn} - z_n)}{var(z_{cn} - z_n) + \frac{2\sigma^2}{I} - \frac{2\sigma^2}{I \cdot C}}
\end{aligned}$$

This expression shows that $\mathbb{E}\hat{b}^{SS} = 0$ if and only if $\beta var(z_{cn} - z_n) = 0$ (no class effects) even when I and C are finite, as with b^{LM}. The degree of attenuation bias is larger than for b^{LM} because class quality is calculated using a smaller number of peer test scores. A similar calculation as above implies that b^{SS} is attenuated relative to β by approximately 35.5%. This estimate roughly matches the 12% difference between the split-sample and leave-out mean estimates in Column 2 of Tables 8a and Appendix Table 6.

4. Peer effects and reflection bias. Throughout the above analyses, we have assumed that there are no peer effects. With peer effects, the assumption $z_{cn} \perp a_{icn}$ does not hold. We expect z_{cn} and a_{icn} to be positively correlated with peer effects as a higher ability student has a positive impact on the class. This leads to an upward bias in both b^{LM} and b^{SS} due to the reflection problem. To characterize the magnitude of this bias, consider a standard linear-in-the-means model of peer effects, in which

$$z_{cn} = t_{cn} + \frac{\theta}{I}\sum_j a_{jcn}$$

with $t_{cn} \perp a_{jcn}$ for all j. Here t_{cn} represents the component of class effects independent of peer effects (e.g., a pure teacher effect). The parameter $\theta > 0$ captures the strength of peer effects. Averaging across classrooms within a school implies that

$$z_n = t_n + \frac{\theta}{IC}\sum_k\sum_j a_{jkn}.$$

In this model, the leave-out mean proxy of class quality is

$$\Delta s_{cn}^{-i} = s_{cn}^{-i} - s_n^i = \frac{IC}{IC-1}(t_{cn} - t_n) + \theta\frac{IC}{IC-1}(a_{cn} - a_n) + \frac{1}{I-1}\sum_{j\neq i} a_{jcn} - \frac{1}{IC-1}\sum_k\sum_{j\neq i} a_{jkn}$$

and the expectation of the coefficient b^{LM} is

$$\begin{aligned}\mathbb{E}\hat{b}^{LM} &= \frac{cov(y_{icn}-y_n, s_{cn}^{-i}-s_n^{-i})}{var(s_{cn}^{-i}-s_n^{-i})} \\ &= \frac{\beta\cdot\left[\frac{IC}{IC-1}var(t_{cn}-t_n)+(\theta+\theta^2)\sigma^2\frac{C-1}{IC-1}\right]+\rho\theta\sigma^2\frac{C-1}{IC-1}}{\frac{(IC)^2}{(IC-1)^2}var(t_{cn}-t_n)+(2\theta+\theta^2)\sigma^2\frac{IC(C-1)}{(IC-1)^2}+\frac{\sigma^2}{I-1}-\frac{\sigma^2}{I\cdot C-1}}\end{aligned}$$

The last term in the numerator is the reflection bias that arises because a high ability student has both high earnings (through ρ) and a positive impact on peers' scores (through θ). Because of this term, we can again obtain $\mathbb{E}\hat{b}^{LM} > 0$ even when $\beta = 0$. This bias occurs iff $\theta > 0$ (i.e., we estimate $b^{LM} > 0$ only if there are peer effects on test scores). This bias is of order $\frac{1}{I}$ since any given student is only one of I students in a class that affects class quality.

Bounding the degree of reflection bias. We use the estimated impact of KG class quality on 8th grade test scores to bound the degree of reflection bias in our estimate of the impact of class quality on earnings. Recall that the reflection bias arises because a high ability student has better long-term outcomes and also has a positive impact on peers' kindergarten test scores. Therefore, the same reflection bias is present when estimating $\hat{b}^{LM}$ using eighth grade test scores as the outcome instead of earnings.

Denote by $\hat{b}_e^{LM}$ the estimated coefficient on Δs_{cn}^{-i} when the outcome y is earnings and $\hat{b}_s^{LM}$ the same coefficient when the outcome y is grade 8 test scores.[42] Similarly, denote by ρ_e and ρ_s the (within class) correlation between individual kindergarten test score and earnings or eighth grade test score. Under our parametric assumptions, these two parameters can be estimated by an OLS regression $y_{icn} = \alpha_{cn} + \rho s_{icn} + \varepsilon_{icn}$ that includes class fixed effects.

To obtain an upper bound on the degree of reflection bias, we make the extreme assumption that the effect of kindergarten class quality on eighth grade test scores ($\hat{b}_s^{LM}$) is due entirely to the reflection bias. If there are no pure class effects ($var(t_{cn}-t_n)=0$) and peers do not affect earnings ($\beta = 0$),

$$\mathbb{E}\hat{b}^{LM} = \frac{\rho\theta}{\frac{1}{1-\frac{1}{I}}+\frac{2\theta+\theta^2}{1-\frac{1}{IC}}} \simeq \frac{\rho\theta}{(1+\theta)^2} \tag{15}$$

Using equation (15) for $\hat{b}_s^{LM}$ and the estimate of $\hat{\rho}_s$, we obtain an estimate of the reflection bias parameter $\frac{\theta}{(1+\theta)^2} = \hat{b}_s^{LM}/\hat{\rho}_s$. Combining this estimate and the estimate $\hat{\rho}_e$, we can then use equation (15) for $\hat{b}_e^{LM}$ to obtain an upper bound on the $\hat{b}_e^{LM}$ that could arise solely from reflection bias.

We implement the bound empirically by estimating the relevant parameters conditional on the vector of parent and student demographics, using regression specifications that parallel those used in column 3 of Table 4a and column 2 of Table 8a. For eighth grade scores, we estimate $\hat{b}_s^{LM} = 0.057$ (se $= 0.029$) and $\rho_s = 0.597$ (se $= 0.012$), and hence

$$\frac{\theta}{(1+\theta)^2} = \frac{0.057}{0.597} = 0.096.$$

For earnings, we estimate $\rho_e = \$93.79$ (se $= \$9.56$) in Table 4a. Hence, if the entire effect of class

[42] The latest test score we have in our data is in grade 8. We find similar results if we use other grades, such as fourth grade test scores.

quality on earnings were due to reflection bias, we would obtain

$$\hat{b}_e^{LM} = \frac{\rho_e \theta}{(1+\theta)^2} = \$93.79 \cdot 0955 = \$8.95 \text{ (se} = \$5.65)$$

where the standard error (se) is computed using the delta method. This upper bound of \$8.95 due to reflection bias is less than 20% of – and significantly different from – the estimate of $\hat{b}_e^{LM} = \$50.61$ (se = \$15.35) in Table 8a. This rejects the null that $\beta = 0$: there must be significant pure class effects or peer effects in earnings. Note that the degree of reflection bias would be smaller in the presence of class quality effects ($\beta > 0$); hence, 20% is an upper bound on the degree of reflection bias in a linear-in-means model of peer effects.

References

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

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

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

4. Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics* 14:1 (1999), 57–67.

5. Angrist, Joshua D. and Alan B. Krueger, "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, American Statistical Association, 13:2 (1995), 225-235.

6. Angrist, Joshua D. and Victor Lavy, "Using Maimonides' Rule to Estimate The Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics,* 114 (1999), 533-575.

7. Bacolod, Marigee P, "Do Alternative Opportunities Matter? The Role of Female Labor Markets in the Decline of Teacher Quality," *Review of Economics and Statistics,* 89:4 (2007), 737-751.

8. Campbell, Frances A., Craig T. Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson, "Early childhood education: Young adult outcomes from the Abecedarian Project," *Applied Developmental Science*, 6 (2002), 42-57.

9. Card, David and Alan B. Krueger, "Does school quality matter? Returns to education and the characteristics of public schools in the United States," *Journal of Political Economy*, 100 (1992), 1-40.

10. Cascio, Elizabeth and Diane W. Schanzenbach, "First in the Class? Age and the Education Production Function," NBER Working Paper No. 13663, 2007.

11. Cilke, James "A Profile of Non-Filers," U.S. Department of the Treasury, Office of Tax Analysis Working Paper No. 78, July, 1998.

12. Corcoran, Sean P., William N. Evans, Robert M. Schwab, "Changing Labor-market Opportunities for Women and the Quality of Teachers, 1957-2000," *American Economic Review,* 94 (2004), 230-235.

13. Cunha, Flavio, and James J. Heckman, "A New Framework for the Analysis of Inequality," *Macroeconomic Dynamics,* 12 (2008), 315-354.

14. Cunha, Flavio, and James J. Heckman, "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, 43 (2008), 738-782.

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

16. Currie, Janet, and Duncan Thomas, "Does Head Start Make a Difference?" *The American Economic Review,* 85 (1995), 341-364.

17. Currie, Janet, and Duncan Thomas, "Early Test Scores, School Quality and SES: Longrun Effects of Wage and Employment Outcomes," *Worker Wellbeing in a Changing Labor Market*, 20 (2001), 103–132.

18. Deming, David, "Early Childhood Intervention and Life-Cycle Skill Development," *American Economic Journal: Applied Economics,* 1 (2009), 111-134**.**

19. Dee, Thomas S., "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics,* 86 (2004), 195-210.

20. Dee, Thomas S., and Martin West, "The Non-Cognitive Returns to Class Size," NBER Working Paper No. 13994, 2008.

21. Dynarski, Susan, Joshua Hyman, and Diane W. Schanzenbach, "Explaining Inequality in Postsecondary Schooling: Experimental Evidence on the Role of Long-Term Investments in Human Capital," manuscript in preparation (2010).

22. Finn, Jeremy D., and Charles M. Achilles, "Answers and Questions About Class Size: A Statewide Experiment," *American Educational Research Journal,* 27 (1990), 557-577.

23. Finn, Jeremy D., Susan B. Gerber, and Jayne Boyd-Zaharias, "Small Classes in the Early Grades, Academic Achievement, and Graduating from High School," *Journal of Educational Psychology,* 97 (2005), 214-223.

24. Finn, Jeremy D., DeWayne Fulton, Jayne Zaharias, and Barbara A. Nye, "Carry-Over Effects of Small Classes," *Peabody Journal of Education,* 67 (1989) 75-84.

25. Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber, "Project STAR and Beyond: Database User's Guide," Lebanon: Heros, inc., 2007.

26. Garces, Eliana, Duncan Thomas and Janet Currie, "Longer-Term Effects of Head Start," *American Economic Review,* 92 (2002), 999-1012.

27. Graham, Bryan, "Identifying Social Interactions Through Conditional Variance Restrictions," *Econometrica,* 76 (2008), 643-660.

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

29. Hanushek, Eric A., "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," *American Economic Review* 61:2 (1971) 280-288.

30. Hanushek, Eric A., "The Failure of Input-Based Schooling Policies." *Economic Journal* 113(1): F64-F98, 2003.

31. Hanushek, Eric A., "Economic Aspects of the Demand for Teacher Quality," prepared for the *Economics of Education Review,* 2010.

32. Hanushek, Eric A., and Lei Zhang, "Quality-consistent estimates of international schooling and skill gradients," *Journal of Human Capital* 3 (2009),107-143.

33. Heckman, James J., Jora Stixrud, and Sergio Urzua, "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behaviors." *Journal of Labor Economics* 24:3 (2006), 411-482.

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

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

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

37. Heckman, James J., and Yona Rubinstein, "The Importance of Noncognitive Skills: Lessons from the GED Testing Program," *American Economic Review*, 91 (2001), 145-149.

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

39. Hoxby, Caroline M., "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 4 (2000), 1239-1285.

40. Hoxby, Caroline M. and Andrew Leigh, "Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States," *American Economic Review,* 94 (2004) 236-240.

41. Internal Revenue Service. *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, 2010.

42. Jacob, Brian A., Lars Lefgren and David Sims, "The Persistence of Teacher-Induced Learning Gains," NBER Working Paper No. 14065, 2008.

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

44. Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica,* 75 (2007), 83-119.

45. Krueger, Alan B, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics,* 114 (1999), 497-532.

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

47. Ludwig, Jens and Douglas L. Miller, "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122 (2007), 159-208.

48. Manski, Charles, "Identification of Exogenous Social Effects: The Reflection Problem," *Review of Economic Studies,* 60 (1993), 531-542.

49. Muennig, Peter, Gretchen Johnson, Jeremy Finn, and Elizabeth Ty Wilde, The Effect of Small Class Sizes on Mortality Through Age 29: Evidence From a Multi-Center Randomized Controlled Trial, unpublished mimeo, 2010.

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

51. Murnane, Richard J., John B. Willett, and Frank Levy, "The Growing Importance of Cognitive Skills in Wage Determination," *The Review of Economics and Statistics,* 77 (1995), 251-266.

52. Neal, Derek A., and William R. Johnson, "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy,* 104 (1996), 869-895.

53. Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges, "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis,* 26 (2004), 237-257.

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

55. Rockoff, Jonah E., "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economics Review,* 94 (2004), 247-252.

56. Rockoff, Jonah E., and Douglas Staiger, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives,* forthcoming, 2010.

57. Sanders, William L., and Sandra P. Horn, "Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research," *Journal of Personnel Evaluation in Education,* 12 (1998), 247-256.

58. Schanzenbach, Diane W., "What Have Researchers Learned From Project STAR?" *Brookings Papers on Education Policy,* (2006), 205-228.

59. Schweinhart, Lawrence J., Jeanne Montie, Zongpin Xiang, William S. Barnett, Clive R. Belfield, and Milagros Nores, "Lifetime effects: The HighScope Perry Preschool Study Through Age 40," Ypsilanti: High/Scope Press, 2005.

60. Schweinhart, Lawrence J., H. V. Barnes and D. P. Weikhart, "Significant Benefits, The High-Scope Perry Pre-School Study Through Age 27," Ypsilanti: High/Scope Press, 1993.

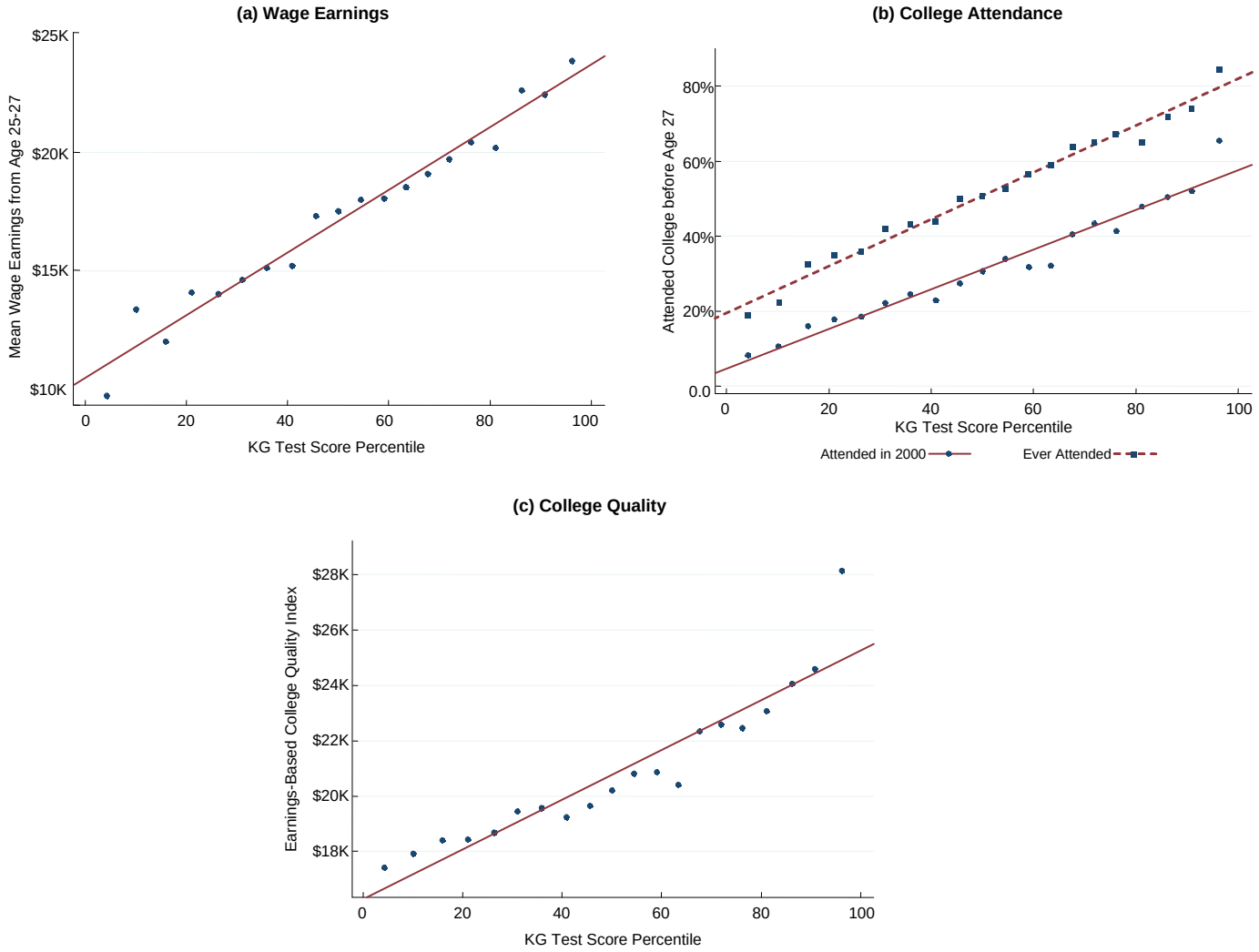
61. Segal, Carmit, "Misbehavior, Education, and Labor Market Outcomes," unpublished mimeo, 2009.

62. Sojourner, Aaron, "Inference on peer effects with missing peer data: Evidence from Project STAR," Carlson School of Management, University of Minnesota mimeo 2009.

63. US Census Bureau. "School Enrollment–Social and Economic Characteristics of Students: October 2008, Detailed Tables," Washington, D.C., 2010 (http://www.census.gov/population/www/socdemo/school.html).

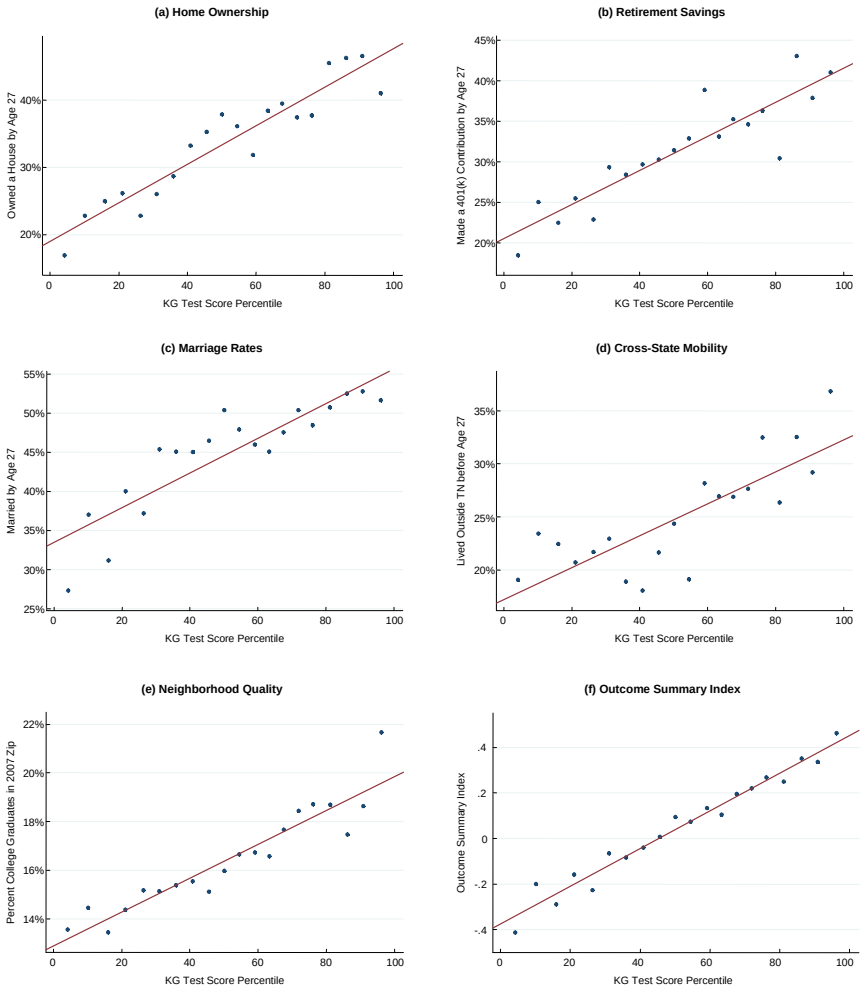
64. Word, Elizabeth., John. Johnston, Helen. P. Bain, B. Dewayne Fulton, Charles M. Achilles, Martha N. Lintz, John Folger, and Carolyn Breda, "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985–1990," Tennessee State Department of Education, 1990.

FIGURE 1
Correlation between KG Test Scores and Adult Outcomes



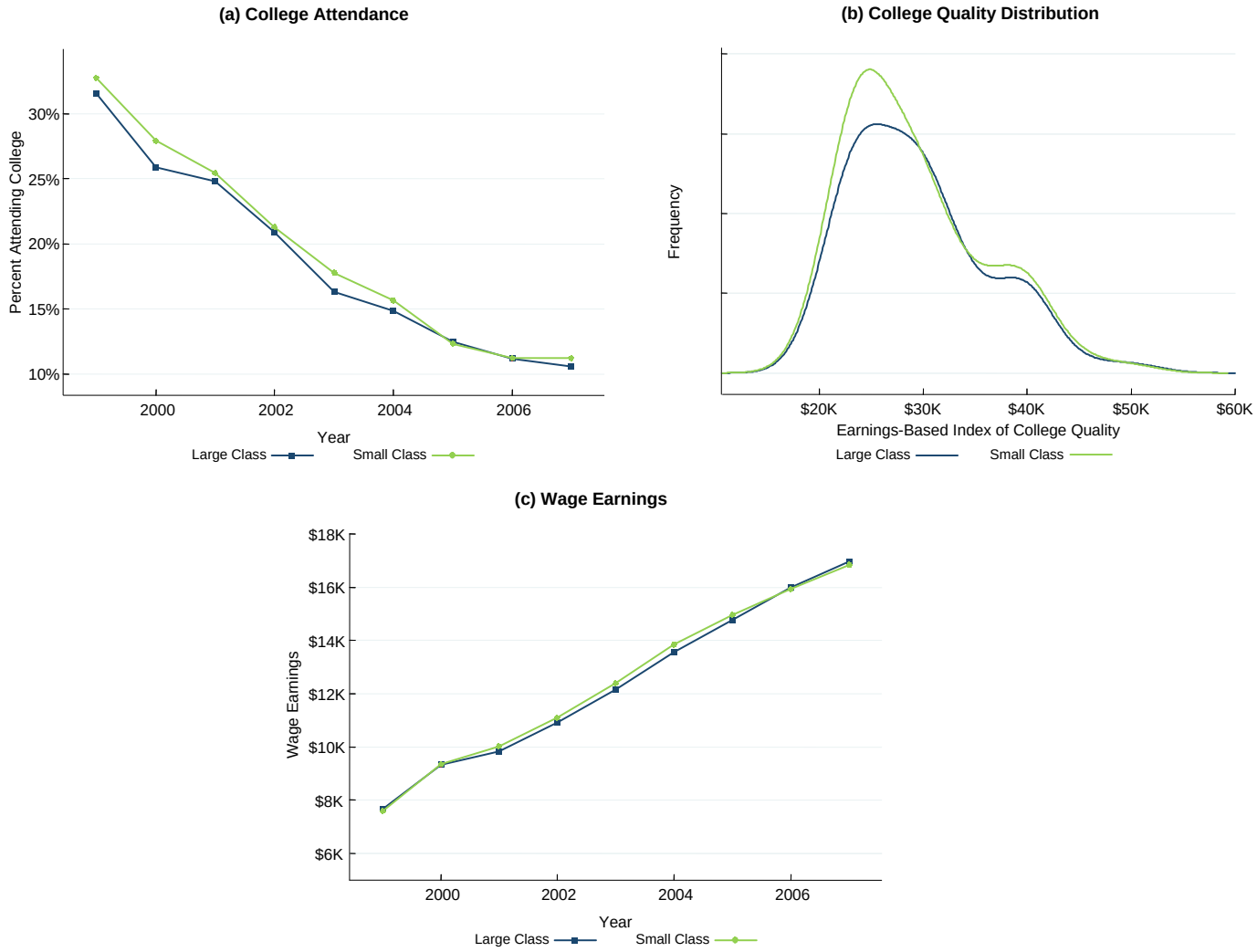
Notes: These figures plot the raw correlations between adult outcomes and kindergarten average test scores in math and reading (measured by within-sample percentile ranks). To construct these figures, we bin test scores into twenty equal sized (5 percentile point) bins and plot the mean of the adult outcome within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. Earnings are mean annual earnings from 2005-2007, measured by wage earnings on W-2 forms; those with no W-2 earnings are coded as zeros. College attendance is measured by receipt of a 1098-T form, issued by higher education institutions to report tuition payments or scholarships, at some point between 1999 and 2007. The earnings-based index of college quality is a measure of the mean earnings of all students who attended each college in the U.S. population at age 28, as described in the text. For individuals who did not attend college, college quality is defined as mean earnings at age 28 of those who did not attend college in the U.S. population. All monetary values expressed in real 2009 dollars.

FIGURE 2
Correlation between KG Test Scores and Other Adult Outcomes



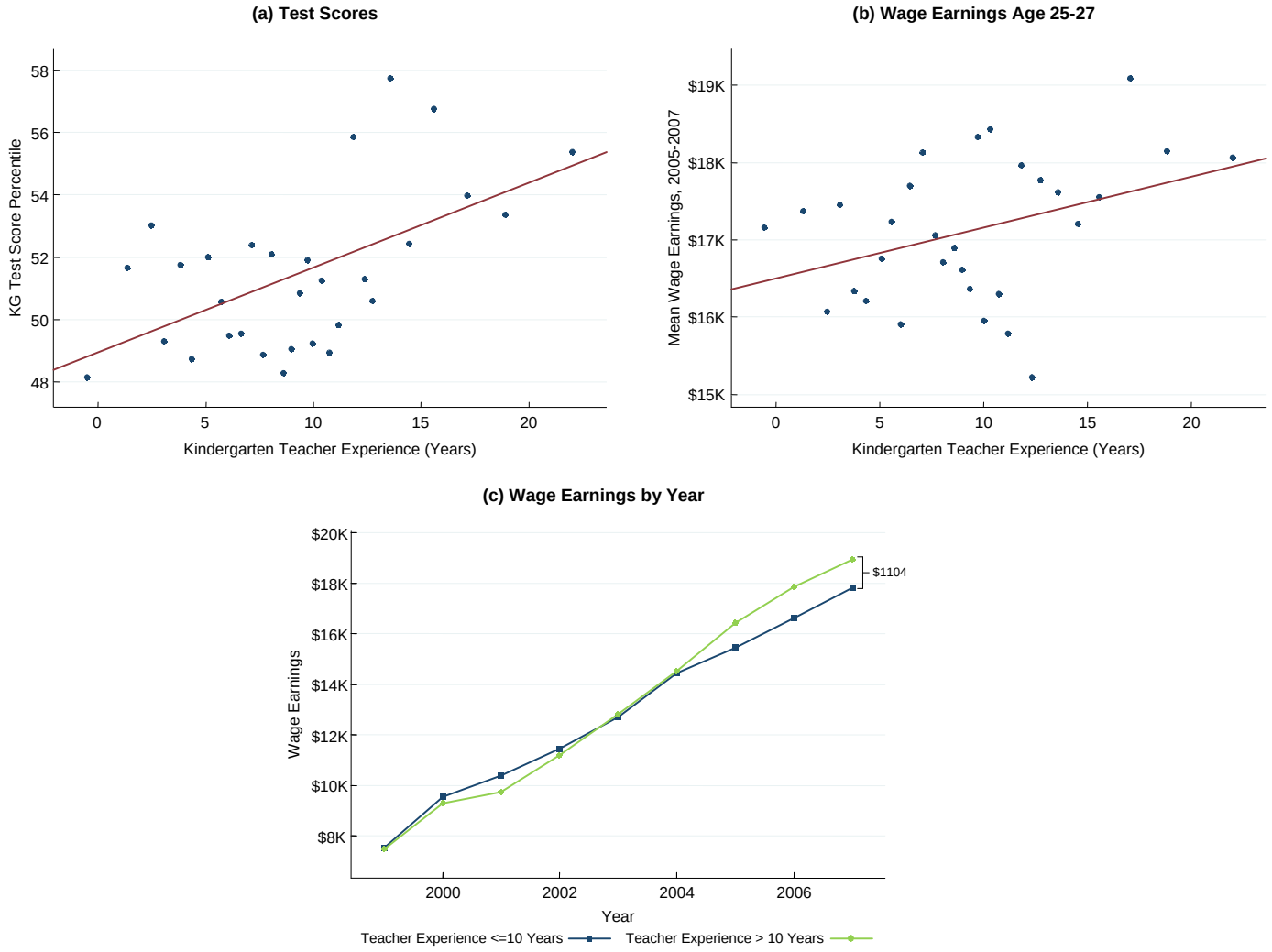
Notes: These figures plot the raw correlations between adult outcomes and kindergarten average test scores in math and reading (measured by within-sample percentile ranks). To construct these figures, we bin test scores into twenty equal sized (5 percentile point) bins and plot the mean of the adult outcome within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. Home ownership is an indicator for whether the individual has a mortgage deduction at any point between 1999 and 2007. 401(k) retirement savings is an indicator for whether the individual makes a 401 (k) contribution between 1999 and 2007. Marital status is measured by whether an individual ever files a joint tax return between 1999 and 2007. ZIP code and state of residence is measured by the address on the 1040 form or the most recent address to which a W-2 form was sent. Cross-state mobility is an indicator for whether the individual ever lived outside TN between 1999 and 2007. Neighborhood quality is defined by the percentage of college graduates in the individual's 2007 ZIP code from the 2000 Census. Summary index is a standardized sum of the five measures in panels (a)-(e). Summary index is scaled such that it has mean 0 and standard deviation of 1.

FIGURE 3
Effects of Class Size



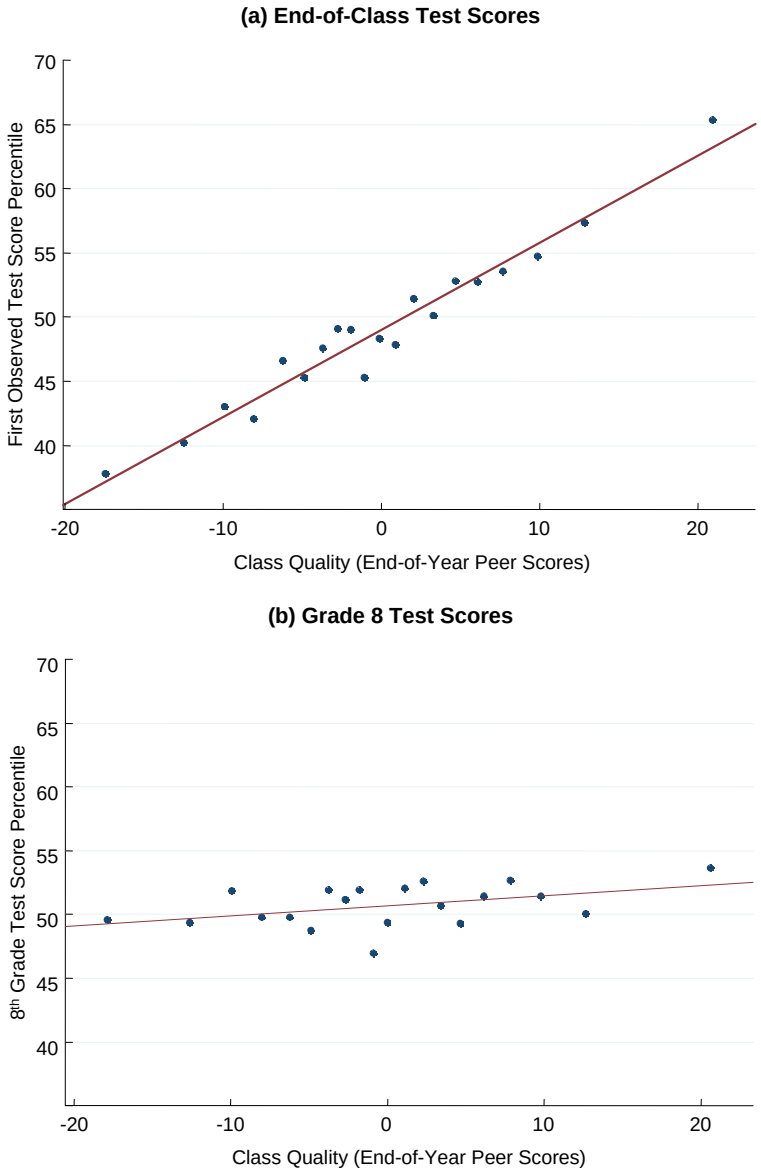
Notes: Panels (a) and (c) show college attendance rates and mean wage earnings by year (from ages 19 to 27) for students in small and large classes. Panel (b) plots the distribution of college quality attended in 2000 using the earnings-based college quality index. Individuals who did not attend college are included in panel (b) with college quality defined as mean earnings in the U.S. population for those who did not attend college. Kernel-smoothed densities in panel (b) are scaled to integrate to total attendance rates for both small and regular classes. All figures adjust for school by entry wave effects to isolate the random variation in class size. In (a) and (c), we adjust for school by wave effects by regressing the outcome variable on school-by-wave dummies and the small class indicator in each tax year. We then construct the two series shown in the figure by requiring that the difference between the two lines equals the regression coefficient on the small class indicator in the corresponding year and the weighted average of the lines equals the sample average in that year. In (b), we compute residual college mean earnings from a regression on school by wave effects and plot the distribution of the residual within small and large classes, adding back the sample mean to facilitate interpretation of units. See notes to Figure 1 for definition of wage earnings and college variables.

FIGURE 4
Effects of Teacher Experience



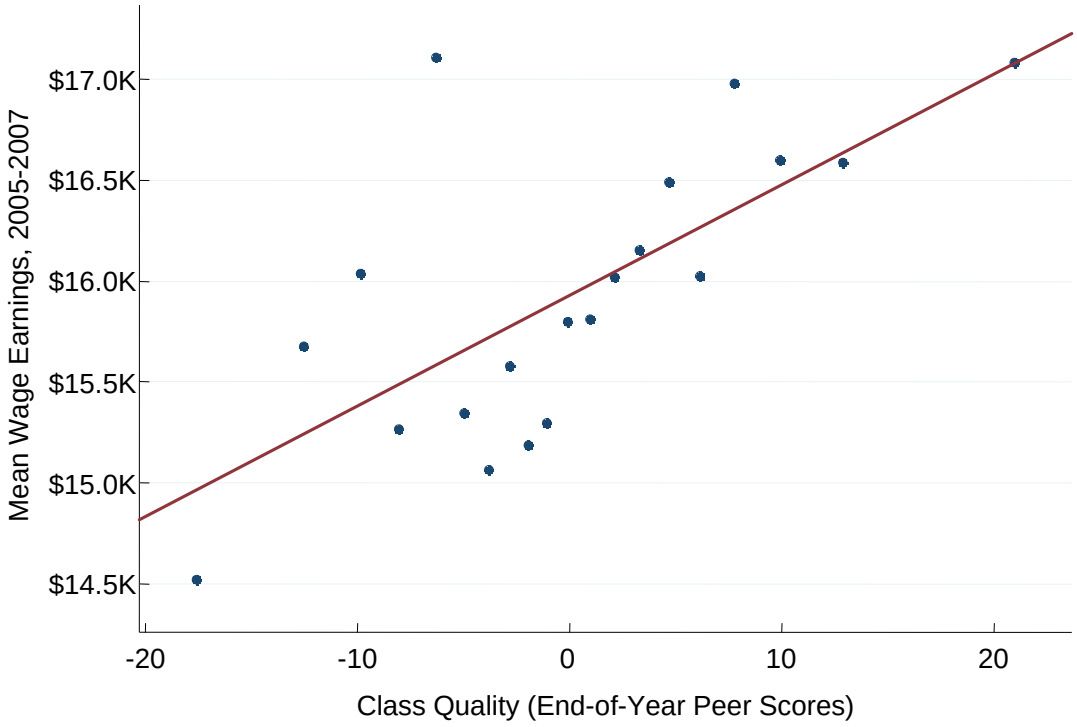
Notes: Panel (a) plots kindergarten average test scores in math and reading (measured by within-sample percentile ranks) vs. kindergarten teacher's years of prior experience. Panel (b) plots mean wage earnings from age 25-27 vs. kindergarten teacher's years of prior experience. Panel (c) plots mean wage earnings by year (from ages 19 to 27) for individuals who had a teacher with less than 10 or more than 10 years of experience in kindergarten. All figures adjust for school by entry wave effects to isolate the random variation in class size. In (a) and (b), we adjust for school by wave effects by regressing both the dependent and independent variables on school-by-wave dummies. We then plot the residuals, adding back the sample means to facilitate interpretation of units. The solid line shows the best linear fit estimated on the underlying data using OLS. In (c), we follow the same procedure used to construct Figure 3c. See notes to Figure 1 for definition of wage earnings.

FIGURE 5
Effects of Class Quality on Test Scores



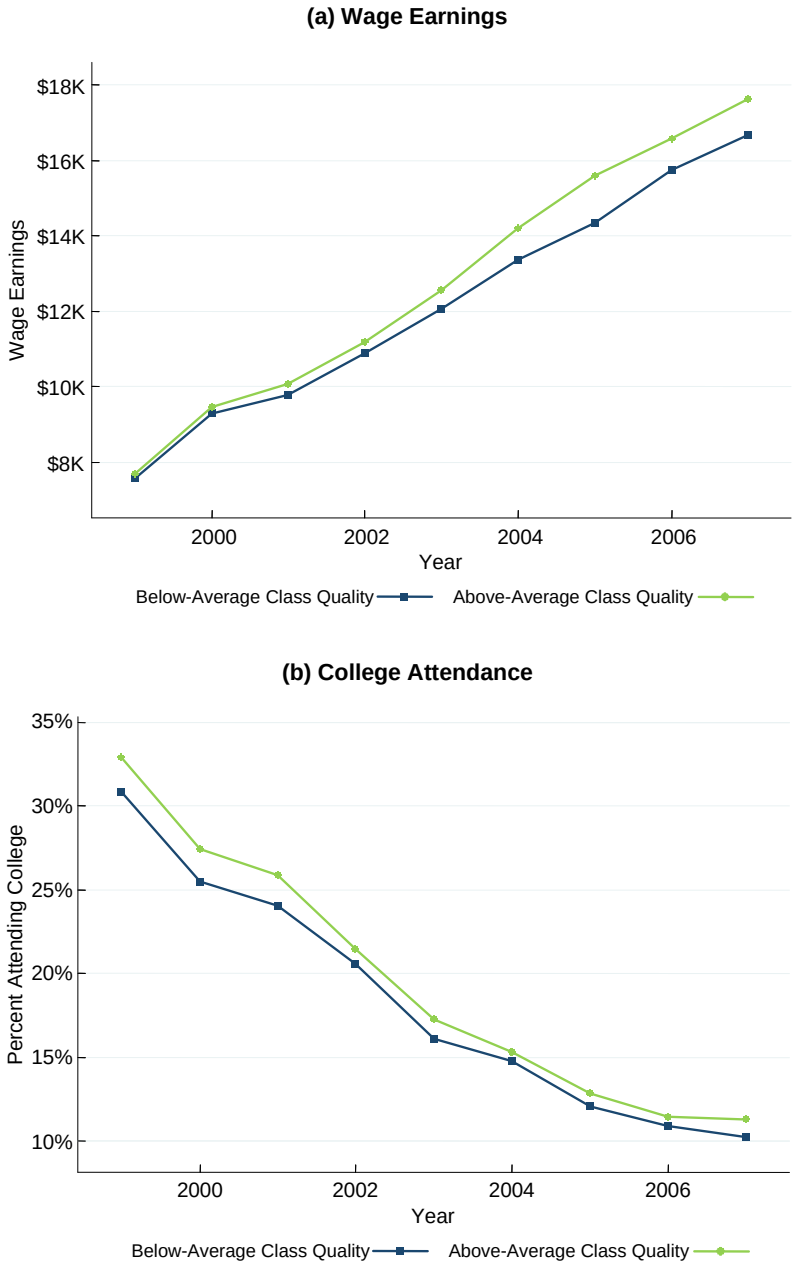
Notes: Panel (a) plots each student's end-of-class test score vs. his class quality, measured as the difference between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e. KG class for KG entrants, 1st grade class for 1st grade entrants, etc.). Correspondingly, end-of-class test score is the student's test score at the end of the first STAR classroom. The coefficient of end-of-class test scores on class quality is 0.68 (se = 0.03), implying that a 1 percentile improvement in class quality causes a 0.68 percentile improvement in test scores. Panel (b) plots the relationship between a student's 8th grade test score and his earliest class quality, defined as in panel (a). The coefficient of 8th grade test scores on class quality is .08 (se = 0.03). Both panels adjust for school by entry wave effects to isolate the random variation in class quality using the technique in Figure 4a. In both panels, we bin test scores into twenty equal sized (5 percentile point) bins and plot the mean of the outcome variable within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS.

FIGURE 6
Effects of Class Quality on Wage Earnings



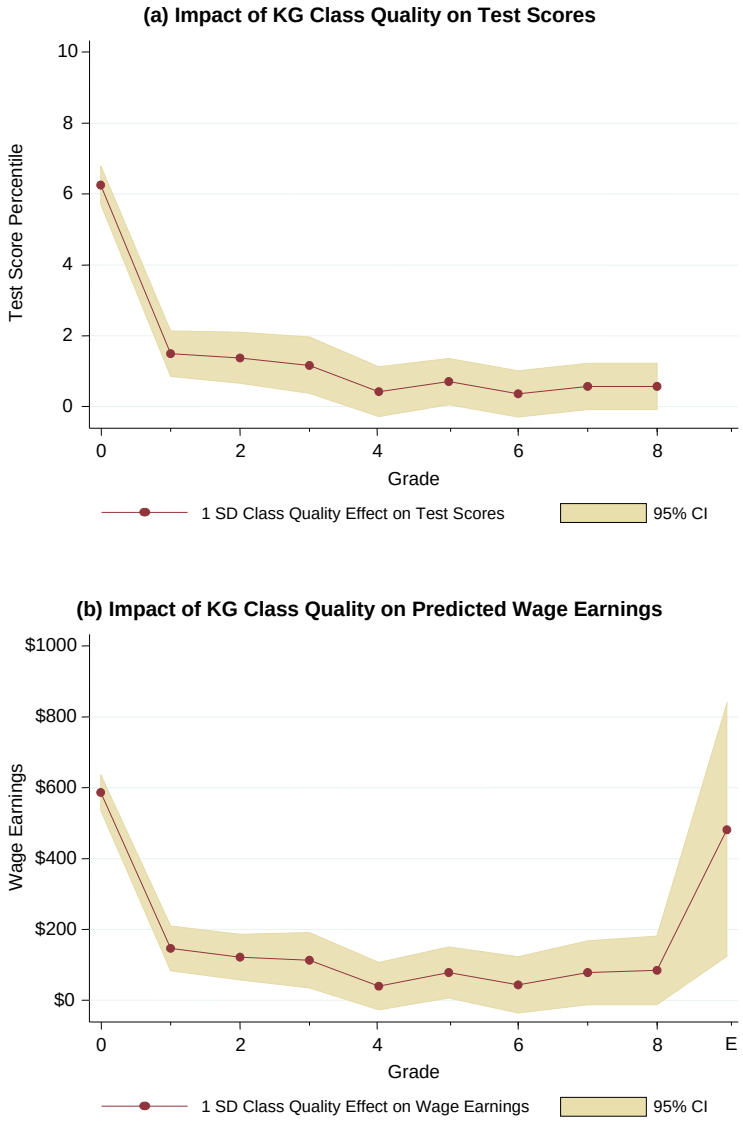
Notes: This figure plots each student's mean wage earnings from 2005-2007 vs. his class quality, measured as the difference between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e. KG class for KG entrants, 1st grade class for 1st grade entrants, etc.). The coefficient of wage earnings on class quality is $57.6 (se = $16.2), implying that a 1 percentile improvement in class quality causes a $57.6 increase in a student's annual earnings. This figure adjusts for school by entry wave effects to isolate the random variation in class quality using the technique in Figure 4a. In this figure, we bin test scores into twenty equal sized (5 percentile point) bins and plot the mean of the outcome variable within each bin. See notes to Figure 1 for definition of wage earnings. The solid line shows the best linear fit estimated on the underlying data using OLS.

FIGURE 7
Effects of Class Quality by Year



Notes: These figures show college attendance rates and mean wage earnings by year (from ages 19 to 27) for students in two groups of classes: those that were above and below the median in terms of quality. Class quality is measured as the difference between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e. KG class for KG entrants, 1st grade class for 1st grade entrants, etc.). Both panels adjust for school by entry wave effects to isolate the random variation in class quality using the procedure in Figure 3c. See notes to Figure 1 for definition of wage earnings and college attendance.

FIGURE 8
Fadeout and Re-Emergence of Class Effects



Notes: Panel (a) shows the impact of a 1 standard deviation improvement in class quality in kindergarten on test scores from grades K-8. Class quality is measured as the difference between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Panel (b) shows the effect of a 1 standard deviation improvement in KG class quality on predicted earnings. It is constructed as follows: first, we run separate regressions of earnings on test scores in each grade, as in column 3 of Table 4a. Next, we estimate the causal impact of KG class quality on test scores for each grade using specifications analogous to column 1 of Table 8a. Finally, we multiply the estimated impact of a 1 SD improvement in KG class quality on test scores by the corresponding OLS regression coefficients of earnings on test score for each grade. The last point in panel (b) shows the actual earnings impact of a 1 SD improvement in KG class quality, estimated using a specification analogous to column 2 of Table 8a. All regressions used to construct these figures are run on the sample of KG entrants and control for school fixed effects and the student and parent demographic characteristics used in Table 8a.

TABLE 1
Summary Statistics

Variable	STAR Sample		U.S. 1979-80 cohort	
	Mean	Standard Deviation	Mean	Standard Deviation
	(1)	(2)	(3)	(4)
Adult Outcomes:				
Average wage earnings (2005-2007)	$15,912	$15,558	$20,500	$19,541
Zero wage earnings (2005-2007)	13.9%	34.5%	15.6%	36.3%
Attended college in 2000 (age 20)	26.4%	44.1%	34.7%	47.6%
Attended college by age 27	45.5%	49.8%	57.1%	49.5%
Owned a house by age 27	30.8%	46.2%	28.4%	45.1%
Made 401 (k) contribution by age 27	28.2%	45.0%	31.0%	46.2%
Married by age 27	43.2%	49.5%	39.8%	48.9%
Moved Out of TN by age 27	27.5%	44.7%		
Percent college graduates in 2007 ZIP code	17.6%	11.7%	24.2%	15.1%
Deceased before 2010	1.70%	12.9%	1.02%	10.1%
Parent Characteristics:				
Average household income (1996-98)	$48,010	$41,622	$65,660	$53,844
Mother's age at child's birth	24.98	6.53	26.29	6.17
Married between 1996 and 2008	64.8%	47.8%	75.7%	42.9%
Owned a house between 1996 and 2008	64.5%	47.8%	53.7%	49.9%
Made a 401 (k) contribution between 1996 and 2008	45.9%	49.8%	50.5%	50.0%
Missing (no parent found)	13.9%	34.6%	23.9%	42.6%
Student Background Variables:				
Female	47.2%	49.9%	48.7%	50.0%
Black	35.9%	48.0%		
Eligible for free or reduced-price lunch	60.3%	48.9%		
Age at kindergarten entry	5.65	0.56		
Kindergarten Teacher Characteristics				
Experience (years)	9.27	5.8		
Post-BA Degree	34.8%	47.6%		
Black	16.2%	36.9%		
Number of Observations	10,992		22,568	

Notes: Adult outcomes, parent characteristics, and student age at KG entry are from 1996-2008 tax data; student background variables (except age) and KG teacher characteristics are from STAR database. Columns 1 and 2 are based on the sample of STAR students who were successfully linked to U.S. tax data. Columns 3 and 4 are based on a 0.25% random sample of the US population born in the same years as the STAR cohort (1979-80). All available variables are defined identically in the STAR and US samples. Earnings are average individual earnings from age 2005-2007, measured by wage earnings on W-2 forms; those with no W-2 earnings are coded as zeros. College attendance is measured by receipt of a 1098-T form, issued by higher education institutions to report tuition payments or scholarships. Home ownership is measured by those who report mortgage interest payments on a 1040 or 1098 tax form. 401(k) contributions are reported on W-2 forms. Marital status is measured by whether an individual files a joint tax return. State and ZIP code of residence are taken from the 1040 form or the most recent address where a W-2 form was sent. Percent college graduates in the student's 2007 ZIP code is based on data on percent college graduates by ZIP code from the 2000 Census. Birth and death information are as recorded by the Social Security Administration. We link STAR participants to their parents by finding the earliest 1040 form from 1996-2008 on which the STAR student is claimed as a dependent. We are unable to link 13.9% of the STAR children (and 23.9% of the U.S. cohort) to their parents; the summary statistics reported for parents exclude these observations. Parent income is average adjusted gross income (AGI) across 1996-98, when STAR participants are aged 16-18. For parents who do not file, household income is defined as zero. Other parent variables are defined in the same way as student variables. Student's age at kindergarten entry is defined as age (in days, divided by 365.25) on Sep. 1 1985, the age at which children in the 1979-80 cohort would normally start KG. Teacher experience is the number of years taught at any school before the current year. All monetary values are expressed in real 2009 dollars.

TABLE 2
Randomization Tests

Dependent Variable:	Test Score	Small Class	Teacher Experience	Teacher Has Post BA Deg.	Teacher is Black	
	(%)	(%)	(Years)	(%)	(%)	p value
	(1)	(2)	(3)	(4)	(5)	(6)
Parent's income ($1000s)	0.085 (0.008) [11.20]	-0.003 (0.014) [-0.249]	-0.003 (0.002) [-1.183]	0.014 (0.017) [0.835]	-0.001 (0.011) [-0.077]	0.848
Mother's age at STAR birth	0.177 (0.039) [4.494]	0.029 (0.072) [0.407]	0.014 (0.011) [1.197]	-0.023 (0.088) [-0.262]	-0.001 (0.056) [-0.017]	0.654
Parents have 401 (k)	1.861 (0.518) [3.594]	1.455 (0.942) [1.544]	0.190 (0.149) [1.269]	0.038 (1.142) [0.033]	-1.039 (0.728) [-1.427]	0.501
Parents own home	0.564 (0.567) [0.995]	-0.007 (1.026) [-0.007]	-0.181 (0.164) [-1.103]	-0.921 (1.257) [-0.733]	-0.244 (0.802) [-0.305]	0.435
Parents married	-0.810 (0.636) [-1.274]	0.803 (1.153) [0.697]	0.272 (0.189) [1.439]	-0.411 (1.443) [-0.285]	-0.127 (0.921) [-0.138]	0.820
Student female	3.509 (0.460) [7.631]	-0.226 (0.832) [-0.271]	0.198 (0.135) [1.466]	0.193 (1.032) [0.187]	-0.455 (0.658) [-0.691]	0.502
Student black	-8.779 (0.920) [-9.546]	0.204 (1.656) [0.123]	0.363 (0.293) [1.237]	2.838 (2.242) [1.266]	1.324 (1.430) [0.926]	0.995
Student free-lunch	-8.112 (0.597) [-13.58]	-0.291 (1.085) [-0.269]	-0.284 (0.174) [-1.635]	0.199 (1.327) [0.150]	-1.012 (0.847) [-1.195]	0.371
Student's age at KG entry	-0.479 (0.484) [-0.990]	-0.828 (0.856) [-0.967]	-0.053 (0.195) [-0.272]	-1.543 (1.486) [-1.039]	-0.348 (0.949) [-0.367]	0.567
p value of F test	0.000	0.261	0.242	0.886	0.718	
Observations	9,939	10,992	6,005	6,005	5,984	

Notes: Columns 1-5 each report estimates from an OLS regression of the dependent variable on the variables listed in the rows of the table and school by entry wave fixed effects. Standard errors are reported in parentheses and t-statistics in square brackets. Test score is the average test score in math and reading (measured by within-sample percentile ranks) in the student's year of entry into a school participating in STAR. Small class is an indicator for assignment to a small class upon entry. Teacher characteristics are for kindergarten teachers. Independent variables are pre-determined parent and student characteristics; see notes to Table 1 for definitions of these variables. P value reported at bottom of columns 1-5 is for an F test of the joint significance of the variables listed in the table. Each row of column 6 reports a p value from a separate OLS regression of the pre-determined variable listed in the corresponding row on school and class fixed effects (omitting one class per school). The p value is for an F test of the joint significance of the class fixed effects. The regressions in column 6 are estimated using the subsample of students who entered in kindergarten. Some observations have missing data on parent characteristics, free-lunch status, race, or mother's age at STAR birth. Columns 1-5 include these observations along with four indicators for missing data on these variables. In column 6, observations with missing data are excluded from the regressions with the corresponding dependent variables.

TABLE 3
Tests for Differential Match and Death Rates

Dependent Variable:	Matched		Deceased	
	(%)	(%)	(%)	(%)
	(1)	(2)	(3)	(4)
Small class	-0.019	0.079	-0.010	-0.006
	(0.476)	(0.418)	(0.293)	(0.293)
p value on F test on class effects	0.951	0.888	0.388	0.382
Demographic controls		x		x
Observations	11,571	11,571	10,992	10,992

Notes: First row of each column reports coefficients from OLS regressions on small class indicator and school by entry wave fixed effects, with standard errors in parentheses. Second row reports a p value from a separate OLS regression of the dependent variable on school and class fixed effects (omitting one class per school). The p value is for an F test of the joint significance of the class fixed effects. Matched is an indicator for whether the STAR student was located in the tax data using the algorithm described in Appendix A. Deceased is an indicator for whether the student died before 2010 as recorded by the Social Security Administration. Columns 1-2 are estimated on the full sample of students in the STAR database; columns 3 and 4 are estimated on the sample of STAR students linked to the tax data. Specifications 2 and 4 control for the following demographic characteristics: student gender, free-lunch status, age, and race and a quartic in the claiming parent's household income interacted with parent's marital status, mother's age at child's birth, whether the parents own a home, and whether the parents make a 401 (k) contribution between 1996 and 2008. Some observations have missing data on parent characteristics, free-lunch status, race, and mother's age at STAR birth; these observations are included along with four indicators for missing data on these variables.

TABLE 4a
Cross-Sectional Correlation between Test Scores and Earnings

Dependent Variable:	Wage Earnings			Log Wage Earnings	Wage Earnings	
	($)	($)	($)	(Log $)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)
KG Test Percentile	131.7	143.0	93.79		-8.439	105.5
	(7.63)	(9.32)	(9.56)		(13.35)	(9.46)
KG Test *z* Score				0.180		
				(0.026)		
8th Grade Test Percentile					156.5	
					(12.00)	
Parental Income Percentile						157.7
						(8.54)
Class Fixed Effects		x	x	x	x	x
Demographic controls			x	x	x	
Adjusted R^2	0.05	0.09	0.17	0.11	0.18	0.16
Observations	5,621	5,621	5,621	5,154	4,202	5,621

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. Sample includes STAR participants who are matched to tax data, entered a STAR school in kindergarten, and have a non-missing kindergarten test score. Dependent variable is mean earnings from 2005-2007 (including zeros for non-participants) except in column 4, where it is the log of mean earnings and zeros are excluded. KG test percentile is the within-sample percentile rank of the student's average score in math and reading. 8th grade test percentile is defined analogously. KG test z score is the mean of the student's scores in math and reading divided by the standard deviations of the within-sample test score distributions. Parental income percentile is the parent's percentile rank in the US population household income distribution. Columns 2-6 include KG class fixed effects to isolate non-experimental variation in test scores. Columns 3-5 control for the vector of demographic characteristics defined in notes to Table 3.

TABLE 4b
Cross-Sectional Correlation Between Test Scores and Other Adult Outcomes

Dep. Var.:	College in 2000	College by Age 27	College Quality	Home Owner	Have 401 K	Married	Moved Out of State	College Grads in 2007 Zip	Summary Index
	(%)	(%)	($)	(%)	(%)	(%)	(%)	(%)	(% of SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KG Test Percentile	0.398	0.527	36.21	0.136	0.100	0.048	0.145	0.053	0.492
	(0.026)	(0.028)	(2.73)	(0.029)	(0.029)	(0.029)	(0.029)	(0.007)	(0.059)
Observations	5,621	5,621	5,621	5,621	5,621	5,621	5,354	5,367	5,621

Notes: Each column reports coefficients from OLS regressions, with standard errors reported in parentheses. Sample includes STAR participants who are matched to tax data, entered a STAR school in kindergarten, and have a non-missing kindergarten test score. KG test percentile is the within-sample percentile rank of the student's average score in math and reading. See notes to Table 1 for definition of dependent variables. The earnings-based index of college quality is a measure of the mean earnings of all the graduates of each college in the U.S. population at age 28, as described in the text. For individuals who did not attend college, college quality is defined by mean earnings at age 28 of those who did not attend college in the U.S. population. All regressions control for class fixed effects and the demographic characteristics defined in notes to Table 3. Summary index is a standardized sum of the five dependent variables in columns 4-8 and is scaled to have mean 0 and standard deviation of 1.

TABLE 5
Effects of Class Size on Adult Outcomes

Dependent Variable:	Test Score	College in 2000	College by Age 27	College Quality	Wage Earnings	Summary Index
	(%)	(%)	(%)	($)	($)	(% of SD)
	(1)	(2)	(3)	(4)	(5)	(6)
Small class (no controls)	4.81	2.02	1.91	119.4	4.09	5.06
	(0.55)	(0.96)	(1.09)	(94.1)	(342.2)	(2.10)
Small class (with controls)	4.76	1.78	1.57	109.1	-124.0	4.61
	(0.54)	(0.87)	(0.99)	(89.3)	(324.5)	(2.01)
Observations	9,939	10,992	10,992	10,992	10,992	10,992
Mean of Dep. Var.	48.67	26.44	45.50	27,115	15,912	0.00

Notes: Each column reports coefficient on small class indicator from two separate OLS regressions, with standard errors in parentheses. Small class indicator is defined as initial assignment to a small class. All specifications include school by entry wave fixed effects to isolate random variation in class assignment. The estimates in the second row (with controls) are from specifications that control for the full vector of demographic characteristics defined in the notes to Table 3. Test score is the average math and reading test score at the end of the year in which the student enters a STAR school (measured in percentiles). Summary index and college quality are defined in notes to Table 4b; remaining dependent variables are defined in notes to Table 1.

TABLE 6
Observable Teacher and Peer Effects

Dependent Variable:	Test Score	Wage Earnings		Test Score	Wage Earnings	
	(%)	($)	($)	(%)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher with >10 years of experience	3.18	1093				
	(0.712)	(453.7)				
Teacher experience (years)			57.13			
			(37.70)			
Teacher has post-BA deg.	-0.848	-261.1	-204.7			
	(0.769)	(489.7)	(493.5)			
% Black classmates				-6.97	-1,757	
				(5.02)	(3,063)	
% Female classmates				9.74	-67.53	
				(2.51)	(1,535)	
% Free lunch classmates				-7.53	-284.6	
				(2.63)	(1,593)	
Classmates' mean age				-3.24	-25.78	
				(2.41)	(1,440)	
Classmates' mean pred. score						-15.95
						(90.65)
Observations	5,601	6,005	6,005	9,939	10,992	10,992

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. All specifications control for school by entry wave fixed effects, an indicator for initial assignment to a small class, and the demographic characteristics defined in Table 3. Test score is the average math and reading test score at the end of the year in which the student enters a STAR school (measured in percentiles). Wage earnings is the individual's mean wage earnings from 2005-2007 (including zeros for non-participants). Specifications 1-3 include only students who enter STAR schools in kindergarten. Teacher experience is the number of years the KG teacher taught at any school before the current year. Classmates' characteristics are defined based on the classroom that the student enters in the first year he is in a STAR school and omit the own student. Classmates' mean predicted score is constructed by regressing test scores on school by entry wave fixed effects and the demographic characteristics defined in notes to Table 3 and then taking the mean of the predicted scores.

TABLE 7
Kindergarten Class Effects: Analysis of Variance

Dependent Variable:	Grade K Scores	Grade 8 Scores	Wage Earnings			
	(1)	(2)	(3)	(4)	(5)	(6)
p value of F test on KG class fixed effects	0.000	0.419	0.047	0.026	0.020	0.042
SD of class effects (RE estimate)	8.77%	0.000%	$1,497	$1,520	$1,703	$1,454
Demographic controls	x	x		x	x	x
Large classes only					x	
Observable class chars.						x
Observations	5,621	4,448	6,025	6,025	4,208	5,983

Notes: First row of each column reports a p value from an OLS regression of the dependent variable on school and class fixed effects (omitting one class per school). The p value is for an F test of the joint significance of the class fixed effects. Second row reports estimated standard deviation of class effects from a model with random class effects and school by entry wave fixed effects. Grade 8 scores are available for students who remained in Tennessee public schools and took the 8th grade standardized test any time between 1990 and 1997. Both KG and 8th grade scores are measured by within-sample percentile ranks. Wage earnings is the individual's mean wage earnings from 2005-2007 (including zeros for people with no wage earnings). All specifications are estimated on the subsample of students who entered a STAR school in kindergarten. All specifications except 3 control for the vector of demographic characteristics defined in notes to Table 3. Column 5 limits the sample to large classes only. This column identifies pure KG class effects because students who were in large classes were re-randomized into different classes after KG. Column 6 replicates 4, adding controls for the following observable classroom characteristics: indicators for small class, above-median teacher experience, black teacher, teacher with degree higher than a BA, and classmates' mean predicted score. Classmates' mean predicted score is constructed by regressing test scores on school by entry wave fixed effects and the demographic characteristics defined in notes to Table 3 and then taking the mean of the predicted scores.

TABLE 8a
Effects of Class Quality on Wage Earnings

Dependent Var: Test Score		Wage Earnings					
	(%)	($)	($)	($)	($)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Class Quality (peer scores)	0.662	50.61	45.79	56.14	53.44	47.70	38.55
	(0.024)	(15.35)	(22.13)	(16.19)	(20.20)	(23.65)	(27.31)
Predicted peer scores				-127.0			
				(95.46)			
Lagged class quality							8.13
							(29.16)
Waves	All	All	All	All	Wave 0	Wave ≥1	Wave ≥1
Large class only			x				
Observable class chars.				x			
Observations	9,939	10,959	8,095	10,859	6,025	4,934	4,839

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. Class quality is measured as the difference (in percentiles) between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e. KG class for KG entrants, 1st grade class for 1st grade entrants, etc.). Test score is the average math and reading test score at the end of the year in which the student enters a STAR school (measured in percentiles). Wage earnings is the individual's mean wage earnings from 2005-2007 (including zeros for people with no wage earnings). All specifications control for school by entry wave fixed effects and the vector of family background characteristics defined in notes to Table 3. Column 3 limits the sample to large classes only. Column 4 replicates 2, adding controls for the observable classroom characteristics defined in notes to Table 7. Column 5 restricts the sample to kindergarten entrants; column 6 includes only those who enter in grades 1-3. Column 7 replicates 6 and adds a control for lagged class quality, measured as the difference between mean test scores of the student's classmates and (grade-specific) schoolmates in the previous year.

TABLE 8b
Effects of Class Quality on Other Adult Outcomes

Dependent Variable:	College in 2000	College by Age 27	College Quality	Summary Index
	(%)	(%)	($)	(% of SD)
	(1)	(2)	(3)	(4)
Class Quality (peer scores)	0.096	0.108	9.328	0.250
	(0.041)	(0.047)	(4.231)	(0.095)
Observations	10,959	10,959	10,959	10,959

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. Class quality is measured as the difference (in percentiles) between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e. KG class for KG entrants, 1st grade class for 1st grade entrants, etc.). Summary index is a standardized sum of five outcomes (home ownership, 401 (k) saving, marital status, cross-state mobility, and percent college graduates in zip code) and is scaled to have mean 0 and standard deviation of 1. College quality is an earnings-based index as defined in notes to Table 4b; college attendance measures are defined in notes to Table 1. All specifications control for school by entry wave fixed effects and the vector of demographic characteristics defined in notes to Table 3.

TABLE 9
Effects of KG Class Quality on Non-Cognitive Skills

Dependent Variable:	Wage Earnings		Grade 8 Math + Reading	Grade 4 Score Math + Reading	Grade 4 Score Non-Cog	Grade 8 Score Math + Reading	Grade 8 Score Non-Cog	Grade 8 Score Math + Reading	Grade 8 Score Non-Cog
	($)	($)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Grade 4 non-cog. score	106.4	87.7	0.059						
	(14.7)	(19.5)	(0.018)						
Grade 4 math + reading score		36.4	0.671						
		(23.3)	(0.021)						
Class Quality (peer scores)				0.047	0.153	0.064	0.128		
				(0.040)	(0.070)	(0.037)	(0.061)		
Teacher with >10 Years Experience								0.292	2.597
								(0.818)	(1.323)
Observations	1,671	1,360	1,254	4,023	1,671	4,448	1,780	4,432	1,772

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. Grades 4 and 8 non-cognitive scores are based on teacher surveys of student behavior across four areas: effort, initiative, interest in class, and disruptive behaviors. We average the four component scores and convert them into within-sample percentile ranks. Math + reading scores are average math and reading test scores (measured in percentiles) at the end of the relevant year. Class quality is measured as the difference (in percentiles) between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates in kindergarten. All specifications include only the subsample of students who entered a STAR school in kindergarten. All specifications control for school by entry wave fixed effects and the vector of demographic characteristics defined in notes to Table 3.

TABLE 10
Impacts of Class Size and Quality on Tax Liabilities

Dpendent Variable:	% Years With Tax Liability > 0			Mean Tax Liability		
	(%)	(%)	(%)	($)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)
Small Class	0.446			19.69		
	(0.755)			(44.50)		
Teacher with >10 Years Experience		-0.797			-53.15	
		(1.021)			(64.63)	
Class Quality (peer scores)			0.084			1.54
			(0.036)			(2.11)
Observations	10,992	6,005	10,959	10,992	6,005	10,959
Mean of Dep. Var.	44.62	48.27	44.62	1,340	1,535	1,342

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. The dependent variable in columns 1-3 is the fraction of years between 1999 and 2007 in which the student reported a positive tax liability. The dependent variables in columns 4-6 is the mean tax liability over the same time period (including zeros for years with no tax liability). All specifications control for school by entry wave effects and the vector of demographic characteristics defined in notes to Table 3. Small class indicator is defined as initial assignment to a small class. See notes to Table 8 for definition of class quality.

APPENDIX TABLE 1
Correlations of Earnings Over the Life Cycle

Age	Correlation between Wage Earnings at Age x and x+6
18	0.36
19	0.36
20	0.37
21	0.41
22	0.47
23	0.55
24	0.60
25	0.62
26	**0.65**
27	0.67
28	0.69
29	0.70
30	0.71
31	0.72
32	0.74
33	0.75
34	0.75
35	0.77
36	0.77
37	0.78
38	0.79
39	0.79
40	0.80
41	0.80
42	0.81
43	0.81
44	0.81
45	0.81
46	0.80
47	0.80
48	0.80
49	0.79
50	0.78

Notes: This table presents correlations between individual mean wage earnings 1999-2001 and individual mean wage earnings 2005-2007 (including zeros for people with no wage earnings) for different ages in a 3% random sample of the US population. Age is defined as age on December 31, 2000. Individuals with mean wage earnings greater than $200,000 in 1999-2001 or 2005-2007 are omitted. Our most common earnings outcome is STAR subject mean wage earnings from 2005-2007. The typical STAR subject was 26 on December 31, 2006. The row in bold implies that STAR subjects' mean wage earnings 2005-2007 are predicted to correlate with their mean wage earnings 2011-2013 (when STAR subjects are approximately aged 31-33) with a coefficient of 0.65.

APPENDIX TABLE 2
Correlation between Test Scores and Adult Outcomes: Heterogeneity Analysis

Dependent Var.:	College in 2000	College by Age 27	College Quality	Wage Earnings	Summary Index
	(%)	(%)	($)	($)	(% of SD)
	(1)	(2)	(3)	(4)	(5)
Blacks	0.355	0.535	32.19	114.9	0.486
	(0.049)	(0.055)	(4.67)	(15.15)	(0.105)
Whites	0.414	0.521	37.64	89.02	0.480
	(0.032)	(0.033)	(3.39)	(12.15)	(0.073)
Males	0.391	0.539	34.92	76.05	0.375
	(0.037)	(0.041)	(3.99)	(14.86)	(0.088)
Females	0.384	0.527	36.21	124.0	0.598
	(0.040)	(0.041)	(3.98)	(12.89)	(0.085)
Free lunch eligible	0.287	0.481	20.47	86.70	0.419
	(0.032)	(0.039)	(2.66)	(11.72)	(0.076)
Not elig. for free lunch	0.551	0.581	57.96	94.83	0.566
	(0.047)	(0.043)	(5.42)	(16.96)	(0.101)

Notes: This table replicates selected specifications in Tables 4a and 4b for various subgroups of students. Each cell shows the coefficient on KG test score in an OLS regression limited to the sub-group defined in the row on the dependent variable defined in the column header. Sample includes STAR participants who are matched to tax data, entered a STAR school in kindergarten, and have a non-missing kindergarten test score. Free lunch is an indicator for whether the student was eligible for free- or reduced-price lunch in kindergarten. All regressions control for class fixed effects and the demographic characteristics defined in notes to Table 3. See notes to Tables 4a and 4b for definition of dependent variables.

APPENDIX TABLE 3
Results for Alternative Measures of Wage Earnings

Dependent Var.:	Positive Mean Earnings	Above Median Earnings	Percentile Earnings	Household Income	2007 Wages
	(%)	(%)	(%)	($)	($)
	(1)	(2)	(3)	(4)	(5)
A. Cross-Sectional Correlations					
KG test score	0.075	0.209	0.158	129.3	110.3
	(0.020)	(0.031)	(0.017)	(13.41)	(11.07)
B. Class Size Impacts					
Small class	0.123	0.482	-0.191	241.5	-263.3
	(0.753)	(1.066)	(0.600)	(446.4)	(374.6)
C. Class Quality Impacts					
Class quality	0.062	0.176	0.098	52.40	45.14
(peer scores)	(0.036)	(0.050)	(0.028)	(21.12)	(17.72)
Mean of Dep. Var.	86.14	50.00	50.00	23,883	16,946

Notes: This table replicates the specifications in Column 3 of Table 4a (in Panel A), Column 4 of Table 5 (in Panel B), and Column 2 of 8a (in Panel C) using alternative measures of wage earnings outcomes. The dependent variables in the five columns are: (1) an indicator for having average wage earnings greater than 0 in the years 2005-2007, (2) an indicator for having average wage earnings between 2005-2007 greater than the sample median ($12,553), (3) the within-sample percentile of a student's average wage earnings, (4) total household income for each student, defined as adjusted gross income adjusted for tax-exempt social security and interest payments, and (5) wage earnings in 2007, winsorized at $100,000 (Column 5). See notes to Tables 4a, 5, and 8a for definitions of regression specifications.

APPENDIX TABLE 4

Impacts of Class Size and Quality on Components of Summary Outcome Index

Dependent Variable:	Home Owner	Have 401 K	Married	Moved Out of State	College Grads in 2007 Zip	Predicted Earnings Summary Index
	(%)	(%)	(%)	(%)	(%)	($)
	(1)	(2)	(3)	(4)	(5)	(6)
A. Class Size						
Small Class	0.712	2.888	1.872	1.105	-0.038	438.92
	(0.980)	(0.986)	(1.020)	(1.014)	(0.245)	(188.51)
Observations	10,992	10,992	10,404	10,268	10,404	10,992
B. Class Quality						
Class Quality (peer scores)	0.080	0.053	0.056	0.029	0.019	20.66
	(0.048)	(0.048)	(0.050)	(0.050)	(0.012)	(8.92)
Observations	10,959	10,959	10,375	10,238	10,375	10,959
Mean of Dependent Var.	30.80	28.18	44.83	27.53	17.60	15,912

Notes: Columns 1-5 decompose the impacts of class size and quality on the summary index into its five constituent components. Each column reports coefficients from an OLS regression, with standard errors in parentheses. Panel A replicates the "with controls" specification of column 6 in Table 5 for each component. Panel B replicates column 4 of Table 8b for each component. See notes to those tables for specification and sample definitions. See notes to Table 1 for definition of dependent variables. Column 6 reports impacts on an alternative "predicted income" summary index. This index is constructed by predicting wage earnings from a regression of mean wage earnings from 2005-2007 on the five dependent variables in columns 1-5. The dependent variable in column 6 is the predicted wage earnings measure.

APPENDIX TABLE 5
Impacts of Class Size and Quality: Heterogeneity Analysis

Dependent Var.:	Test Score	College in 2000	College by Age 27	College Quality	Wage Earnings	Summary Index
	(%)	(%)	(%)	($)	($)	(% of SD)
	(1)	(2)	(3)	(4)	(5)	(6)
A. Effect of Small Class						
Blacks	6.871	2.722	5.312	249.0	250.0	6.308
	(0.899)	(1.466)	(1.755)	(139.4)	(493.9)	(3.243)
Whites	3.699	1.065	-0.177	38.94	-348.1	4.388
	(0.674)	(1.101)	(1.201)	(116.0)	(422.9)	(2.581)
Males	4.883	2.594	2.279	244.5	798.3	10.89
	(0.745)	(1.157)	(1.349)	(121.4)	(484.3)	(2.845)
Females	4.360	0.716	0.454	6.638	-1130	-2.599
	(0.798)	(1.359)	(1.491)	(135.9)	(436.5)	(2.932)
Free lunch eligible	5.767	0.837	3.908	-2.517	-251.7	3.162
	(0.710)	(1.013)	(1.302)	(79.10)	(388.7)	(2.529)
Not elig. for free lunch	3.376	3.592	-0.914	296.6	293.0	7.292
	(0.843)	(1.615)	(1.565)	(192.5)	(585.5)	(3.444)
B. Effect of Class Quality (peer scores)						
Blacks	0.732	0.081	0.089	3.197	36.22	0.236
	(0.036)	(0.062)	(0.074)	(5.844)	(20.70)	(0.136)
Whites	0.582	0.069	0.075	12.16	68.17	0.180
	(0.036)	(0.060)	(0.065)	(6.296)	(22.92)	(0.140)
Males	0.654	0.016	0.067	8.257	66.03	0.357
	(0.036)	(0.057)	(0.067)	(5.987)	(23.86)	(0.140)
Females	0.654	0.151	0.131	7.845	32.37	0.032
	(0.037)	(0.065)	(0.072)	(6.541)	(21.00)	(0.141)
Free lunch eligible	0.650	0.058	0.089	2.319	53.14	0.077
	(0.032)	(0.047)	(0.061)	(3.685)	(18.08)	(0.118)
Not elig. for free lunch	0.652	0.149	0.104	22.34	47.11	0.442
	(0.043)	(0.085)	(0.083)	(10.19)	(30.97)	(0.182)

Notes: This table replicates selected specifications in Tables 5 and 8 for various subgroups of students and dependent variables. Panel A replicates the "with controls" specification of Table 5. Panel B replicates the specification of Table 8b. Each cell shows the coefficient on an indicator for small class (Panel A) or class quality (Panel B) in an OLS regression limited to the sub-group defined in the row on the dependent variable defined in the column header. Free lunch is an indicator for whether the student was eligible for free- or reduced-price lunch in kindergarten. See notes to Tables 5 and 8b for definitions of regression specifications and notes to Tables 1 and 2 for definitions of dependent variables.

APPENDIX TABLE 6
Split Sample Measure of Class Quality

Dependent Var:	Test Score	Wage Earnings	College in 2000	College by Age 27	College Quality	Summary Index
	(%)	($)	(%)	(%)	($)	(% of SD)
	(1)	(2)	(3)	(4)	(5)	(6)
Class Quality (peer scores)	0.524	44.86	0.105	0.098	8.46	0.292
	(0.022)	(13.96)	(0.038)	(0.042)	(3.84)	(0.086)
Demographic controls	x	x	x	x	x	x
Observations	9,939	10,959	10,959	10,959	10,959	10,959

Notes: This table replicates columns 1 and column 2 of Table 8a and all columns of Table 8b using the split-sample definition of class quality described in part 3 of Appendix B. Each column reports coefficients from an OLS regression, with standard errors in parentheses. In this table, classrooms are randomly divided into two sub-class groups and class quality for students in each group is measured using mean peer scores in the other group. Each regression specification controls for school by entry wave by sub-class group fixed effects as well as the vector of family background characteristics defined in notes to Table 3. All other aspects of the regression specifications are identical to Table 8a and Table 8b; see notes to Table 8a Table 8b for details.

APPENDIX TABLE 7
Effects of Class Quality on Components of Non-Cognitive Measures

A. Cross-Sectional Correlations

Dep Var.:	Wage Earnings							
	Grade 4				Grade 8			
	($)	($)	($)	($)	($)	($)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Effort	98.31				110.85			
	(19.88)				(15.95)			
Initiative		66.15				73.27		
		(20.01)				(15.37)		
Value			66.65				90.83	
			(18.05)				(15.91)	
Participation				37.36				59.86
				(16.72)				(15.54)

B. Class Quality Impacts

Dep Var.:	Grade 4				Grade 8			
	Effort	Initiative	Value	Particip	Effort	Initiative	Value	Particip
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class Quality (peer scores)	0.144	0.157	0.093	0.122	0.145	0.072	0.128	0.173
	(0.069)	(0.069)	(0.070)	(0.073)	(0.070)	(0.071)	(0.069)	(0.069)

Notes: This table decomposes the impacts on the non-cognitive measures shown in Table 9 into their four constituent components: effort, initiative, value, and participation. These four non-cognitive measures are constructed from a series of questions intended to measure, respectively, student effort in class, initiative, whether a student perceives school/class as "valuable", and participation in class. Each of the four variables is measured by within-sample percentile ranks. Panel A shows the cross-sectional association between each non-cognitive measure and wage earnings using an OLS regression that controls for test scores in the same grade. Panel A replicates the regression specification in column 2 of Table 9. In columns 1-4 of Panel A, the independent variables are non-cognitive measures in grade 4; in columns 5-8, they are the same measures in grade 8. Panel B shows the impact of class quality on each non-cognitive measure, replicating the specification in column 4 of Table 9. Standard errors in parentheses. See notes to Table 9 for definitions of the regression specifications.