Test Scores, Schools, and the Geography of Economic Opportunity

Sulagna Mookerjee
Georgetown University SFS-Qatar

David Slichter
Binghamton University (SUNY)*

January 2019

Abstract

Do standardized test scores in a community indicate whether schools there are effective at producing human capital? Counties with high average test scores produce high-earning adults. But, using data from North Carolina, we find that counties' *effects* on test scores are either uncorrelated (for low-income kids) or *negatively* correlated (for high-income kids) with their effects on income in adulthood. We argue with a simple model that this is probably because the inputs directly responsible for counties' effects on test scores do not substantially increase income. In particular, we directly demonstrate that differences in test score production have little to do with teacher quality. Our results suggest that differences in test score production across places are not necessarily a useful measure of the quality of schools.

JEL classification: I24, J24, J62
Keywords: Human capital, intergenerational mobility, value-added

*Email: sulagna.mookerjee@georgetown.edu and slichter@binghamton.edu. Thank you to our advisors Gregorio Caetano and Josh Kinsler for detailed comments. We are also grateful to Minjeong Kim, Plamen Nikolov, Amy Schwartz, and participants at the Stockman Conference, the Canadian Economic Association, the North America Summer Meeting of the Econometric Society, and Binghamton University for helpful comments. Thank you to Kara Bonneau and NCERDC for furnishing additional technical details. We thank Kate Gerontianos for research assistance. All remaining errors are our own.

1 Introduction

Policy-makers, researchers, and journalists routinely interpret high standardized test scores in a region as evidence that schools there are preparing students for the workforce. This interpretation implicitly rests on two assumptions: (i) differences in test scores across regions are driven by differences in school quality, and (ii) school systems which produce higher test scores are producing better workers.

Are these assumptions correct? In this paper, we argue that, at the unit of analysis of a county, the answer is likely "no."

Using data from North Carolina, we begin by measuring the effects of living in a given county on students' test scores. We confirm our measurements by showing that students who move between counties show the expected change in test score performance.

Next, we grant the premise that differences in counties' effects on test scores are due to schools, and measure whether counties which improve test scores also improve incomes in adulthood for children who are raised there. We find that, for children from low-income families, there is no correlation between a county's effects on test scores and its effects on incomes. (The r-squared of this regression is 0.0004.) For children from high-income families, the correlation is *negative* and significant.

We then turn to understanding why there is not a positive correlation between the effects of a county on test scores and its effects on income in adulthood. We consider three possible explanations. One possibility is lack of statistical power: perhaps test score production is valuable, but counties do not differ much in their production of test scores. A second is that the lack of positive correlation is due to confounding variables. A third possibility is that the inputs which lead counties to produce different test scores do not appreciably increase income in adulthood.

To assess these three explanations, we construct a simple model of the effects of place on incomes through test scores. Using the model, we argue that, for high-income kids in particular, the data cannot plausibly be rationalized with the first two explanations alone. We conclude that the inputs which account for county differences in test score production do not substantially improve incomes in adulthood – at least for high-income kids.

This finding is puzzling if counties' differences in test score production are due to the school system, since teachers are widely believed to be the most important input provided by schools, and teacher quality has been shown to substantially affect income in adulthood (Chetty et al. 2011, Chetty et al. 2014b). We resolve this puzzle by observing teachers who move between counties. If variation in counties' test score effects were solely due to teacher quality, teachers would keep the same measured test score value-added as they move across counties. Instead, teachers who move to higher (lower) value-added counties experience a sudden increase (decrease) in their measured value-added at the time of their move. The magnitude of this jump in measured value-added suggests that little of the variation in counties' effects on test scores is due to teacher quality.

Additional evidence suggests that school-level inputs are also unlikely to be important, as teachers do not change value-added when moving between schools within the same county. (This is also consistent with prior studies arguing for the unbiasedness of value-added measures.) School districts might be quite important, as test score differences emerge close to school district boundaries, but we cannot precisely disentangle the role of district inputs from cultural or environmental amenities. Whatever the exact origins of county differences in test score production are, though, it is unlikely both that school systems are responsible for most of the variation in test scores across counties, and that the school system inputs which are producing higher test scores also produce substantially higher incomes in adulthood.

Collectively, our results suggest that test score production in a county is not necessarily an informative measure of the quality of schools in that county, if school quality is defined in terms of effects on adult incomes.

Additionally, as a secondary finding related to the literature on teacher evaluation, our results suggest that value-added regressions which are unbiased when comparing teachers in the same community (as previously demonstrated by Kane and Staiger 2008, Kane et al. 2013, and Chetty et al. 2014a, and re-confirmed with our data) are likely biased when comparing teachers in different places. This finding may be of interest to policy-makers constructing statewide teacher evaluation systems.

In Section 7, we briefly discuss why counties' test score production might not be helpful for incomes in adulthood. One possibility is that counties' differences in test score production are mostly due to substitution between useful inputs, some of which affect test scores and others of which do not. Another possibility is that short-run gains in numeracy and literacy do not translate into meaningful long-run gains. Finally, literacy and numeracy skills (as distinguished from broader reasoning skills) may be so widespread that they do not command a substantial premium in the labor market. We leave a more detailed exploration to future research.

Two papers perform analysis closely related to our research topic. The first is Chetty and Hendren (2018b), who show that test score levels in a county (adjusted for average household income) are positively correlated with the effects of that county on incomes in adulthood. Our analysis differs from theirs because we measure *effects* of counties on test scores, rather than *levels* of test scores.[1] We replicate a positive correlation of counties' test score levels with counties' effects on income in North Carolina, and interpret the fact that Chetty and Hendren's results differ from ours as an indication that the distinction between levels and effects matters. Combined with their findings, our results suggest that test score levels might proxy for non-school characteristics of a community – but we also cannot rule out the possibility that schools are better in counties with higher test score levels, though only in ways which do not manifest themselves in test scores.

The other most closely related paper to ours is by Rothstein (2018), who argues that the education system is unlikely to explain the geographic pattern of economic

[1]Sufficiently detailed data to isolate place effects on test scores are not available nationally.

opportunity because of the low correlation between intergenerational mobility in a region and the extent to which test scores of high- and low-income students diverge over schooling years in that region. We make two contributions strengthening Rothstein's conclusions. First, we draw similar conclusions from an independent method. In particular, our method can rule out various potential objections to Rothstein's work. Second, our findings help clarify the mechanism behind Rothstein's findings. Section 7 contains a more detailed discussion of the relationship between the papers.

Our finding does not guarantee that test scores are not an informative measure of school quality at other levels of regional aggregation or in other contexts. Prior work (Hanushek and Woessmann 2012, Schoellman 2012) has found evidence suggesting that nations with high test scores also have better schools. Similarly, *within* counties, our results suggest that differences in test score production are driven by teacher quality, suggesting that highly local differences in test score levels might be informative about the quality of education being provided. Finally, our findings do not mean that school quality does not matter or does not vary at the county level, since school quality may not manifest itself in test score performance. Card and Krueger (1992) argue that school quality varies by state, and Biasi (2018a) argues at the unit of analysis of counties that school financing can help explain the geographic pattern of opportunity measured by Chetty et al. (2014c).

The rest of the paper proceeds as follows. Section 2 describes the data. Section 3 presents the econometric model used to study the effects of counties on test scores, and shows evidence that this model captures causation. Section 4 shows correlations of test score effects with test score levels and income effects. Section 5 builds a simple model of effects of test scores on incomes. Section 6 explores which school inputs can explain variation in test score effects. Section 7 discusses implications of the results. Section 8 concludes.

2 Data

We use two sources of data. The first is data on all students who attended a North Carolina public school in grades 3-8, provided by the North Carolina Education Research Data Center (NCERDC). The main standardized test score measure we use is performance on the End-of-Grade (EOG) exams, which are used by the state of North Carolina as the primary measure of student performance. Other key variables include socio-economic and demographic characteristics such as gender, race, and whether the students are enrolled in a free or reduced price lunch program. Students and teachers can be linked longitudinally using unique identifiers.[2] Our primary analysis is based on data from the years 1999-2006, since these are the only years in which data on free/reduced price lunch eligibility is available, but we perform some

[2]The process of linking students to teachers is imperfect, as described in documentation on the NCERDC website (https://childandfamilypolicy.duke.edu/research/nc-education-data-center/). See Appendix D for further discussion of issues related to linking teachers.

robustness checks using data from 1994-1998 and 2007-2013 as well.

Our second source of data are estimates of the effects of growing up in a given county on income in adulthood, produced by Chetty and Hendren (2018b) and available on their project website.[3] In addition to estimates of county effects on incomes, their public data contains a variety of county covariates used in their analysis. We refer the reader to their paper and its companion (Chetty and Hendren 2018a) for a complete description of how their estimates are constructed.

Descriptive statistics for the North Carolina data are shown in Table 1. We use free/reduced price lunch eligibility as a proxy for whether a student is from an above- or below-median income household.[4] The table shows that somewhat under half of students are in fact from free/reduced price lunch households.[5]

Table 1: Descriptive statistics

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
# Schools	1,452	1,458	1,480	895	802	796
Proportion male	0.513	0.511	0.510	0.512	0.511	0.508
Proportion white	0.579	0.587	0.592	0.593	0.600	0.608
Proportion FRL	0.445	0.435	0.427	0.415	0.395	0.371
# County switchers		38,101	41,992	57,187	43,968	41,407
# Students	838,714	832,033	833,689	839,016	833,732	815,851

FRL stands for free/reduced price lunch eligible. County switchers are students who were observed in another county in North Carolina the previous year.

A potential concern related to internal validity is that we only use data from the public school system, while test score comparisons across places do not necessarily

[3]The URL is http://www.equality-of-opportunity.org/.

[4]The free/reduced price lunch program offers subsidized meals to students with income levels below 185% of the federal poverty line. Eligibility is commonly used as a proxy for family income.

[5]In addition to the fact that not exactly half of students are free/reduced price lunch eligible, the declining fraction of free/reduced price lunch students in older grades is a reminder that parental income is dynamic, and that a snapshot measure of parental income (free/reduced price lunch status eligibility is based on contemporaneous parental income) may not correspond with the measure of parental income used by Chetty and Hendren. However, our results are unlikely to be sensitive to this measurement error, since the variance of county effects on test scores is similar for above- and below-median households, and since there is a substantial correlation (.76) between our measure of test score effects for high- and low-income students. This suggests that incorrect categorization of parental income is unlikely to drive our results.

only use students in public schools, and Chetty and Hendren's estimates are not restricted to children in public schools. However, according to the 5-Percent Public Use Microdata Sample from the 2000 Census, 89.7% of children above the age of 5 and below the age of 18 in North Carolina were currently enrolled in a public school. This fraction rises to 92.2% among those children in that age range who are currently enrolled in any school. Even among children in that age range from households with above-median income, 89.6% of currently enrolled students attend a public school.[6]

3 Econometric model

We begin by measuring the average effect of living in each county in North Carolina on test scores for high- and low-income students. To do this, we use a regression model with a set of controls which has been found to approximately eliminate bias due to student unobservables in models of teacher value-added to test scores (Kane and Staiger 2008, Kane et al. 2013, Chetty et al. 2014a).

For each grade-year (e.g. 5th graders in 2001), we use ordinary least squares regression to estimate the following equation using data only from that grade-year:

$$A_{ij} = \alpha^{gt} + X_i'\beta^{gt} + \gamma_j^{gt} + u_{ij},$$

where i indexes students, j indexes teachers, gt indexes parameters by the grade-year to which the regression is restricted, A is normalized test score performance, X is a vector of covariates including a cubic of lagged test score performance, race/ethnicity, gender, and free/reduced price lunch status, γ_j is a teacher fixed effect, and u is a mean-zero error term. Year and grade subscripts are not included on the student variables to reflect that each regression is estimated using cross-sectional data from a single grade only. The teacher fixed effect is included so that the parameters α^{gt} and β^{gt} will be identified only using within-classroom variation, producing the same coefficients on the controls X as would be produced in the sort of teacher value-added models which guide our choice of controls. We estimate the model variously using three measures of test score performance: math score, reading score, and total score, where total score is produced by summing the normalized math and reading scores, then normalizing this sum.

Next, using the resulting parameter estimates (denoted with hats), we construct the difference between each student's actual test score performance and their expected test score performance based on their characteristics X, and label this difference for student i in grade g in year t as $studentVA_{ijgt}$:

$$studentVA_{ijgt} := A_{ijgt} - \left[\widehat{\alpha^{gt}} + X_{igt}'\widehat{\beta^{gt}}\right].$$

[6] "Median" here is defined in terms of household income among those children in the specified age range living in North Carolina.

Here, we subscript student variables with gt to denote that this constructs a panel dataset of $studentVA$. Note that $studentVA_{ijgt}$ is an estimate of $\gamma_j^{gt} + u_{ijgt}$, i.e. it captures the influence of teacher quality as well as any other factors affecting achievement, possibly including non-teacher inputs which vary by county.

We treat $studentVA_{ijgt}$ as an estimate of the effect of the inputs received by student i in year t on i's test score achievement. This estimate may be noisy at an individual level. However, prior research suggests that, with large enough samples, averages of $studentVA$ for students within the same classroom provide an approximately forecast-unbiased measure of teacher value-added to test scores.[7]

Analogously to such teacher value-added models, to estimate the effects of living in county c on test scores in grade g and time t, we simply take the sample average of $studentVA_{ijgt}$ for all i in grade g living in county c at time t. Our key results also use separate estimates of counties' effects on test scores for high- and low-income children, which are constructed by further restricting this sample average to students of the appropriate parental income type, measured using free/reduced price lunch eligibility. Finally, we construct estimates of the effects of living in county c on test scores in grade g as the average of our estimates for that county in that grade among all available years.[8] In the remainder of the paper, we use $countyVA_{cg}^{P}$ to refer to this average for students of parental income type P in grade g in county c.

Causal interpretation We will interpret the average $studentVA$ within a county as reflecting the effect of living in that county on test score achievement. However, a natural concern with value-added regressions of this kind is whether they capture the effects of counties on student achievement, or instead reflect differences across counties in student unobservables (e.g. Rothstein 2017).

The sort of student unobservables which would lead our measure to be biased (if correlated with county of residence) are those features of the student which would influence their test score no matter which county the student lived in. This may include characteristics such as innate ability or parent quality.

Recall that the main claim of this paper is that it is not jointly true that (i) test score differences across places can be attributed to schools, and (ii) the school inputs generating these test score differences also prepare students for the labor force. If the county value-added estimates were driven by differences in student unobservables such as those described above, this would suggest (i) is incorrect, supporting our main conclusion. Therefore, for the purposes of this paper, it is perhaps more conservative

[7] With small samples, it is common to shrink estimates in the direction of zero to preserve forecast-unbiasedness. See Chetty et al. (2014a).

[8] For the years 2003, 2004, and 2005, the data does not distinguish between students who are not eligible for free/reduced lunches and students who attend schools that do not report participation in the FRL program, for whom the FRL status is therefore not applicable. Our main analysis (Tables 4 and 5) therefore excludes the observations from these years. We include these years for other components of the analysis in order to maximize sample size, but our results are robust to dropping these years.

to err on the side of interpreting the county value-added estimates as approximately causal.[9] Nonetheless, it is also possible to gather some direct evidence on whether this is the case, and it is helpful for interpreting our subsequent results.

To assess whether unobservables are important, we study the change in students' test score achievement residuals $studentVA$ (the averages of which we would like to interpret as the effects of place, including indirectly through teacher quality) as students move across counties. If important student unobservables are present, we would expect that students' achievement residuals would be less sensitive to their context than our econometric model suggests. On the other hand, if our estimates are causal, then students who switch counties should experience a growth in their achievement equal to the growth predicted by the model. More precisely, the average change in achievement will equal the model-predicted gain whenever the model delivers forecast-unbiased estimates of county effects on test scores.[10]

Note that students may also have additional stability in their measured $studentVA$ if they sort into similar quality amenities, such as teachers or schools, in different counties. Observing students who move between counties would therefore understate the differences in average amenity quality across counties. In Appendix A, we show that the change in school and classroom quality that students experience as they move between counties is, on average, slightly smaller than the change in county quality.

While the student and family traits which lead students to sort to similar amenities in different counties fit the definition of a "student unobservable" above, we would like to incorporate all differences in average teacher and school quality (not simply the differences which are not eliminated by sorting) in our measure of counties' effects. This is because, for example, if student sorting led students to attend identical quality schools no matter where they lived, yet places differ in their average school quality, we believe most people would still take this to mean that school quality differs across places. This indicates that the definition of county effects on test

[9]In principle, this would not be conservative if the set of controls in the value-added regression absorbed counties' effects on test scores. (In this scenario, counties would have effects on test scores which were possibly attributable to schools, while at the same time our measure would be driven by student unobservables.) However, we are not aware of evidence that this set of controls tends to absorb a large fraction of variation in contemporaneous school inputs' effects on test score performance. Furthermore, the main message of our paper is that test scores cannot be used to directly measure the quality of a county's schools. If a value-added model does not recover the effects of local schools on test scores, we view it as unlikely that this information can be reliably recovered from currently available data. If local test scores to some extent reflect the effects of schools, but these effects cannot be detected with standard econometric methods, test scores are for all intents and purposes uninformative.

[10]See Chetty et al. (2014a) for the definition of forecast-unbiasedness. Note that not all forecast-unbiased estimators yield exactly the full causal effect. For example, an estimator which assigns a zero causal effect on test scores to all counties (relative to a randomly selected other county) would be forecast-unbiased. If our estimates are forecast-unbiased, then the variance of true causal effects is either as high as we measure, or even higher. This would suggest, for example, that the variance of place effects on test scores measured in Section 5 is smaller than the true variance.

scores which is best aligned with our research question should not remove variation in school quality that is eliminated through sorting, and an ideal test of the causal model should account for this sorting. Therefore, our test of the regression model is a test of whether, among students who move across counties, $studentVA$ responds to moving as much as the model predicts based on the change in classroom-level residuals.

Our test proceeds as follows. Among students who switch counties (say, from county c' and classroom j' in year $t-1$ to county c and classroom j in year t), we estimate the effect of the change in an individual student's achievement residual on the change in their estimated classroom value-added using the following regression:

$$\Delta studentVA_{it} = \alpha_0 + \alpha_1 \Delta classroomVA_{it} + v_{it},$$

where, letting $studentVA_{it}$ be the student residual of student i in year t and $classroomVA_{jt}$ be the average student residual (excluding i) in the classroom j of student i during year t,

$$\Delta studentVA_{it} := studentVA_{it} - studentVA_{it-1}$$

and

$$\Delta classroomVA_{it} := classroomVA_{jt} - classroomVA_{j't-1}.$$

Due to the small number of students in each class, classroom value-added is imprecisely estimated in any given year, which would lead to attenuation bias in the estimate of α_1.[11] To correct this, we use the standard approach of instrumenting for a mismeasured variable using an independent noisy measurement of the same variable (see Wooldridge 2015).[12] Specifically, we instrument for $\Delta classroomVA_{it}$ using the difference in classroom value-added reversing the years, i.e.

$$\Delta Rev\_classroomVA_{it} := classroomVA_{jt-1} - classroomVA_{j't}.$$

Results, broken down by subject (subscript M for math and R for reading), are shown in Table 2.

If our value-added measure were biased due to student unobservables (not including tendency to sort into similar quality classrooms), we would expect to see a coefficient smaller than 1 in these regressions. Instead, we see a coefficient just larger than 1 for math and just smaller than 1 for reading, with both estimates statistically indistinguishable from 1. This suggests that our econometric model is capturing the effects of local amenities on test scores, rather than simply capturing student unobservables.

[11] Additionally, if there are peer effects, classroom value-added may be influenced by the presence of i.

[12] When estimating a single teacher's value-added, a standard approach to handling sampling error in value-added is to shrink estimates in the direction of zero (e.g. Chetty et al. 2014a). In our case, though, determining the number to which to shrink $\Delta classroomVA$ requires additional assumptions that are not necessary for the instrumental variables solution.

Table 2: Change in student residuals

	$\Delta StudVA_M$	$\Delta StudVA_R$
$\Delta ClassVA_M$	1.081**	
	(0.062)	
$\Delta ClassVA_R$		0.894**
		(0.147)
First Stage		
$\Delta Rev\_ClassVA_M$	0.527**	
	(0.012)	
$\Delta Rev\_ClassVA_R$		0.329**
		(0.013)
N	5,129	5,069

** indicates $p < 0.01$. Robust standard errors in parentheses. The bottom panel shows the first stage. The top panel shows the instrumental variable results. To minimize measurement error in teacher assignment, results are for grade 4 and 5 classrooms with at least 10 and no more than 40 students.

It would be possible to coincidentally recover a coefficient of 1 in this regression if both our value-added measure *and* the change in student residuals at the time of a move were (equally) biased due to unobservables. Therefore, in Appendix B, we perform placebo tests for bias in the student-switching quasi-experiment. We find that the location to which a student moves is not correlated with abnormal changes in residuals before or after the move.

Summarizing, in order *not* to believe that the estimates from the value-added model are approximately forecast-unbiased, one would need to believe that (i) unobservables bias the value-added model, (ii) the student-switching quasi-experiment also suffers from bias, (iii) the magnitude of bias in these designs is coincidentally equal, yet (iv) the bias in the student-switching experiment manifests itself only exactly at the time of the move, while the bias in the value-added model is, presumably, permanent. Furthermore, from the previous literature and our results in Section 6, one must additionally believe that (v) the unobservables which bias the value-added model across counties do not vary enough within counties to substantially bias local comparisons of teacher value-added. We feel it is more likely that the value-added model is approximately forecast-unbiased.

It is important to note the word "approximately," though. Each estimate among our robustness checks comes with sampling error, so we cannot guarantee exact forecast-unbiasedness. In fact, because our value-added estimates are not experimental, we consider it more likely than not that there are some modest remaining differences in student unobservables across counties which are simply too small to meaningfully affect our robustness checks. We will argue in Section 4 that, if we have failed to fully control for differences in student unobservables, that would most likely bias our main results upwards. That is, the correlation between counties' *true*

effects on test scores and on incomes would be *more negative* than our main results imply.

4 Correlation of test score and income effects

We next turn to measuring whether counties which are good at producing test scores are also good at producing incomes in adulthood.

These estimates are analogous to exercises measuring the returns to high value-added teachers (Chetty et al. 2011, Chetty et al. 2014b), which treat a teacher's value-added as the measure of the intervention received, even though teachers have many characteristics and abilities – some of which may be correlated with that teacher's effects on test scores (Jackson 2017) and may help account for the value of high value-added teachers. Similarly, we begin simply by asking whether high value-added counties are beneficial. This can help assess whether test score performance is a useful indicator of which counties are good places to raise children, but does not directly tell us the value of test score production per se.

We use regressions to measure the relationship between our estimates of county effects on test scores and Chetty and Hendren's estimates of county effects on incomes. Chetty and Hendren offer multiple estimates, available on their project website.[13] We primarily use their preferred estimates of place on percentile rank in the adult household income distribution, since these estimates seem to be the least noisy.[14] Chetty and Hendren offer separate estimates for children at the 25th and 75th percentiles of parental income. We denote Chetty and Hendren's estimates for the effects on income percentile at age 26 of growing up in county c for children at the 25th percentile of parental income as Inc_c^0, and at the 75th percentile of parental income as Inc_c^1.

Our data spans grades 3-8 and the estimation of value-added requires lagged test scores, so we are able to estimate test score effects in grades 4-8. To aggregate these estimates, we construct a test score measure

$$Test_c^P := \sum_{g=4}^{8} countyVA_{cg}^P,$$

which is defined for $P \in \{0, 1\}$ denoting parental income type. We will use $P = 1$ to denote high-income, i.e. not free/reduced price lunch eligible. We then estimate the regression

$$Inc_c^P = \gamma_0^P + \gamma_1^P Test_c^P + \eta_c^P$$

[13] See http://www.equality-of-opportunity.org/data/.

[14] In their data, the key variables are *pct_causal_p25_kr26* and *pct_causal_p75_kr26* from the data files for Online Table 2.

for $P = \{0, 1\}$, where η is a mean-zero error term. The results are shown in Columns 1 and 3 of Table 3. Panel A uses Chetty and Hendren's preferred measure of household income percentile; as a robustness check, Panel B uses place effects on individual income percentiles instead.[15] There is no correlation at all between income and test score effects of a county for children from low-income families; the r-squared of this regression is .0004. Surprisingly, for children from high-income families, there is a *negative* correlation which is statistically significant at the 1% level.

Table 3: Income effects on test score effects

	(1) Inc_c^0	(2) Inc_c^0	(3) Inc_c^1	(4) Inc_c^1
	Panel A: Household Level			
$Test_c^0$	.039	.077		
	(.198)	(.146)		
$Test_c^1$			$-.235^{**}$	$-.249^{**}$
			(.085)	(.076)
Controls	N	Y	N	Y
R^2	.0004	.5582	.0760	.2936
N	96	96	94	94
	Panel B: Individual Level			
$Test_c^0$	.083	.120		
	(.116)	(.120)		
$Test_c^1$			$-.327^{**}$	$-.355^{**}$
			(.122)	(.138)
Controls	N	Y	N	Y
R^2	.0059	.1131	.0707	.0735
N	96	96	94	94

** indicates $p < .01$. Robust standard errors in parentheses. Controls in Column 2 are the fractions of adults in the $1^{st}, 6^{th}$ and 7^{th} deciles of the income distribution, and controls in Column 4 are the fractions of adults in the 6^{th} and 7^{th} deciles of the income distribution for Panel A and in the 1^{st} decile in Panel B.

In this section, we focus solely on significance, but we will turn to interpreting the magnitudes of these estimates with the help of a model in Section 5.

[15]The rate of return to teacher value-added measured in Chetty et al. (2014b), which we use later in this paper, uses a mixture of household and individual income sources which is not exactly analogous to either of the measures used by Chetty and Hendren.

Correlations with test score levels Chetty and Hendren find a positive correlation between test score levels in a county and its effect on incomes.[16] This seems to conflict with our findings in Table 3. But we find a different pattern of correlation for test score *levels* than for test score *effects*, suggesting that the distinction between levels and effects is important and drives the difference between our results and theirs.

Let $Level_c$ be the average normalized total test score performance of 8th grade students in county c, and let $Level_c^P$ be the same average only for students of parental income type P.[17] We regress Inc_c^P on $Level_c$, then on $Level_c^P$, to provide counterparts to Chetty and Hendren's results (which do not divide test score levels by parental income) and to Table 3. The results are shown in Table 4.

Table 4: Income effects on test score levels

	(1) Inc_c^0	(2) Inc_c^0	(3) Inc_c^1	(4) Inc_c^1
$Level_c$	.640**		-.037	
	(.100)		(.070)	
$Level_c^0$		.996**		
		(.117)		
$Level_c^1$				-.040
				(.082)
R^2	.2599	.4767	.0036	.0030
N	99	99	99	99

** indicates $p < .01$. Robust standard errors in parentheses.

For low-income children, test score levels are strongly positively correlated with income production (Columns 1 and 2). For high-income children, there is no significant relationship. Both of these findings differ from the correlations of income production with test score production in Table 3. These results are robust to using test score levels in other grades and to controlling for household income.

We interpret this as evidence that the distinction between test score levels and test score effects is important. Consistent with Reardon (2018), we find an imperfect, though substantially non-zero, correlation between test score levels in a county and the effects of living in that county on test scores. The correlation between $Level_c^0$

[16]Their measure of test scores is the component of average test score performance which is orthogonal to average household income. To an approximation, this is a measure of test score levels, not test score effects of a place.

[17]The average test score is constructed by taking the average normalized total test score in each year individually, then averaging across years.

and $Test_c^0$ is .30; the correlation between $Level_c^1$ and $Test_c^1$ is .68. These correlations are slightly lower when using test score levels in grades prior to 8th grade.

Explaining the lack of positive correlation The lack of a positive correlation between counties' effects on test scores and on incomes is perhaps surprising. While it is intuitive that sorting would account for a large part of differences in test score *levels* across places, differences in *effects* on test scores seem more likely to be driven by schools. If schools are responsible, it is very likely that teacher quality in particular plays an important role. And prior research (Chetty et al. 2011, Chetty et al. 2014b) suggests that teacher quality increases incomes in adulthood.

We devote most of the remainder of the paper to considering three broad explanations for this surprising lack of a positive correlation. The first possible explanation is that differences across counties in test score production are so small that this correlational exercise simply lacks power to detect a signal. A second explanation is that the correlational results in Table 3 are biased, in the sense that they do not capture the return to the amenities which are directly responsible for differences in test score production. A third explanation is that there is not a positive correlation because the amenities which produce test scores are simply not useful for producing incomes. This implies in particular that test score production differences are not driven by teacher quality. We will return to the third possibility only after considering the first two and finding them to be inadequate to explain all of our findings.

To assess the first possible explanation, we will need a model, which we construct in Section 5.

The second explanation – that this (absence of) correlation does not reflect causation – could take two forms: bias from omitted variables and bias due to measurement error.

In order to consider bias due to omitted variables, first we must be more precise about what variables, exactly, would count as "omitted." We will take "omitted variables" to mean community characteristics which (a) do not narrowly, directly produce test scores and (b) which affect incomes in adulthood. For example, average income in a community might be an omitted variable, since average income itself does not produce test scores (though it may affect test scores indirectly through provision of other amenities), and may also affect incomes in adulthood. Note that this definition allows for a negative effect of interest, i.e. after accounting for all omitted variables, since the inputs which are directly involved in producing the test scores may reduce production of untested skills (e.g. if those inputs are teaching to the test in schools, or any form of substitution away from time expenditure on other activities).

The theoretical case for negative omitted variables bias is not obvious ex ante. In fact, there are many theoretical reasons to believe that communities which produce high test scores should on average be *stronger* on other dimensions. For example, if differences in test score effects arise in part because of differences in willingness to pay by communities for amenities related to human capital development, we would

expect counties with higher willingness to pay for the test score amenity to have a higher willingness to pay for non-test score amenities related to human capital development. A negative correlation between $Test_c^P$ and η_c^P would require both that this mechanism is relatively unimportant and that some other mechanism generating omitted variables bias is important.

As a preliminary exploration of possible omitted variables bias, we include additional controls in the above regression to capture broad community characteristics which are not obviously directly implicated in test score production. We would like to control for the full collection of county characteristics used by Chetty and Hendren, which includes variables related to community characteristics like segregation and social cohesion.[18] (However, we do not control for their test score measure or class size measure, since these variables would not count as omitted under our definition of an omitted variable.) This list of characteristics is quite long; there are 44, in addition to 10 variables capturing the fraction of the adult population in each income decile. Since this is large relative to the number of counties in North Carolina, we use the post-LASSO double machine learning method from Belloni et al. (2014) to reduce the dimension of controls.

County effects on test scores turn out not to be strongly correlated with the list of county characteristics provided by Chetty and Hendren. The double LASSO procedure does not select any county characteristic as a predictor of test score effects, either for low-income or high-income students. As an additional measure, then, we instead select characteristics which LASSO finds are both correlated with Chetty and Hendren's measure of test score performance in a national sample, and correlated with effects on income in the North Carolina sample. For low-income students, this yields three controls: the fraction of adults in the 1st, 6th, and 7th deciles of the national income distribution. For high-income students, this yields two controls: the fraction of adults in the 6th and 7th deciles of the national income distribution.

The results are shown in Columns 2 and 4 of Table 3. As expected from the lack of meaningful correlation with test score effects of places, the inclusion of these controls does not appreciably alter the key coefficients, though the explanatory power of the regression increases substantially. This is most consistent with a limited role for omitted variables bias.

Because we do not observe all other features of communities, we cannot rule out the possibility that the regression results would be sensitive to controls which we do not observe. However, in Section 5, we will argue using a simple model that the required degree of omitted variables bias is quite unlikely, at least for high-income kids.

Another potential source of bias in our regressions is measurement error. One source of measurement error is that both the independent and dependent variables are estimated. Our estimates of $Test$ are quite precise, since county value-added is estimated using many students; there are an average of 31,661 observations of

[18] See their Online Table 4 for the full list of characteristics.

$studentVA$ per county, yielding a precision in the sample average akin to estimating an individual teacher's value-added from on the order of 1,000 years of data.[19] As a result, classic attenuation bias due to sampling error is unlikely to drive the result.[20] However, Chetty and Hendren lack sample size to precisely estimate effects of community of residence on incomes using only their preferred identification strategy, which is based on families who move. Therefore their estimates at the level of individual counties are noisy. Of course, error in the left-hand side variable will introduce bias only if it is correlated with the right-hand side variable. Mere sampling error in their data would not produce such correlation, though it might lead to larger standard errors in our regression.

There might also be concerns about bias in Chetty and Hendren's estimates. In response to the small number of movers, to minimize forecast error, their estimates are constructed using a mixture of families who move and outcomes of permanent residents. While the use of permanent county residents may lead to bias in Chetty and Hendren's estimates of county effects on income, it does not necessarily bias the expected coefficients in the regression above. This is because increasing the causal effect of a place on income will generally have the same effect on income both for movers and for permanent residents. Any bias introduced into our regression results through the use of permanent residents would be due to a correlation between test score effects and the characteristics of permanent residents. The regressions with controls described above suggest that this compositional effect is unlikely to explain our results, since observed characteristics of county residents do not seem to be meaningfully correlated with our measure of test score effects. We cannot rule out bias due to a correlation with unobserved permanent resident characteristics, however.

It is also possible that, despite the evidence shown in Section 3, our estimates of test score effects are to some extent biased by student unobservables. This would presumably bias the coefficients of Table 3 in the direction of the coefficients in Table 4 – only strengthening the puzzle of why we do not measure positive correlations between Inc and $Test$.[21]

[19]The smallest county has 2,338 observations. All but two counties have at least 5,000 observations. Even at the parental income type-subject-grade-year level, which is substantially less aggregated than our test score measure used in Table 3, only roughly 5% of the variation in estimated $countyVA$ is due to sampling error.

[20]Another illustration of this point is that our results in Section 3 suggest that our county value-added measure is (approximately) forecast-unbiased, which means that the measurement error in the independent variable is (approximately) uncorrelated with its actual value – a scenario which does not produce the usual attenuation bias.

[21]Another reason to suspect that bias would most likely work in this direction is that individuals' incomes and test score performance are positively correlated (e.g. Murnane et al. 2000, Lazear 2003, Chetty et al. 2011, and Chetty et al. 2014b, among many others), suggesting that student characteristics which increase test score performance usually increase incomes, and would therefore lead to higher values of Chetty and Hendren's estimates, especially since these are influenced by the outcomes of permanent residents. To believe that student unobservables account for the zero and negative correlations we find, one would need to believe (i) the string of beliefs described at the

A further, perhaps more serious, concern is that the cohorts in our data are not the same as those in Chetty and Hendren's data. They use birth cohorts from 1980 to 1986, while our preferred estimates use children in 4th through 8th grade between 2000 and 2006, which presumably corresponds with children born between about 1986 and 1996.[22] If place effects on test scores vary greatly from year to year, this would attenuate our estimates in the direction of 0.

One piece of evidence that this does not drive our results is that the findings are robust to using earlier years of data, though we do not prefer these estimates because the data lacks some observations and variables. We are able to construct estimates of county effects on test scores in some years – 1995, 1998, 1999, and 2000 – in which Chetty and Hendren's cohort would have been in the observed grades.[23] These estimates in earlier years cannot be subdivided into estimates for high- and low-income children, as we do not have a measure of free/reduced price lunch status in these years. (For the same reason, we also cannot control for parental income in the value-added regression.) County test score effects for all students in these earlier years are nearly as correlated with test score effects for low-income (correlation of .64) and for high-income (.59) in our preferred years as the effects for low- and high-income kids are with each other (.76) in those preferred years, suggesting that test score effects are substantially, though not perfectly, stable across time. Not surprisingly, then, the correlations of counties' effects on test scores and on incomes are robust to replacing $Test_c^P$ with county test score effects for all students measured only in these earlier years, with a positive but insignificant ($p = .97$) correlation between overall test score effects and income effects for low-income kids, and a negative (though insignificant, $p = .13$) relationship for high-income kids.[24] Using 1995 data alone, in which our observations align almost perfectly with the birth cohorts used by Chetty and Hendren, produces a negative but insignificant relationship for low-

conclusion of Section 3 which justify a belief in substantial bias of our value-added estimates and (ii) these student unobservables have zero or negative influence on income. Our main results suggest that there are at least *some* inputs which increase test scores without influencing incomes – but we do not view it as obviously more likely that student unobservables should have this negative or zero effect than it is that the inputs responsible for (true) county value-added would. Therefore, we think (ii) is at least no more plausible a priori than our main results; combined with the unlikeliness of (i), we believe our interpretation is more reasonable. If one instead believes that our value-added model is approximately but not exactly forecast-unbiased, then there is no requirement to believe (i) to believe that student unobservables drive our results, but there would be a requirement to replace (ii) with the belief that these student unobservables have extreme negative impacts on income in adulthood, which we view as quite unlikely.

[22]Testing occurs in the spring semester. We label years according to the year in which the spring semester occurs.

[23]We also have data from 1996 and 1997 but exclude these years due to the very small number of non-missing values for key variables in 1996. Missing data in 1996 prevents us from estimating value-added in 1997. We are also not able to estimate value-added regressions residuals for approximately half of students in 1995.

[24]Using effects on individual income as our outcome, the estimates are negative and insignificant both for low-income kids ($p = .69$) and high-income kids ($p = .12$).

income children ($p = .17$) and a negative and significant correlation for high-income children ($p = .02$), though we do not prefer these estimates for the reasons that a large number of student observations are missing and we do not have information on free/reduced price lunch status.

Two additional pieces of evidence suggest to us that, while it surely somewhat affects the estimates, the mismatch of years does not *drive* our results. First, the mismatch of years should shrink our estimates in the direction of zero. Yet the correlation for high-income kids is still significant, and, as we will argue in Section 5, of a substantial magnitude; the oddity is simply that the estimate is negative. Second, as described above, raw test score levels in our preferred years are still strongly correlated with counties' effects on incomes for low-income children.[25]

5 Model

We next construct a simple model of the effects of county of residence on income in adulthood through standardized test scores. This model is used to assess the plausibility of possible explanations for the lack of positive correlation between counties' effects on test scores and their effects on incomes in adulthood.

Our fundamental argument in this section is that the variance of counties' effects on test scores is so large that arguments based on lack of statistical power or omitted variables bias cannot explain the results of the previous section if counties' effects on test scores are due to a productive input like teacher quality. For high-income kids in particular, the data can only be rationalized if the inputs responsible for counties' test score effects do not substantially increase income in adulthood.

The outline of the model description is as follows. First, we define some parameters of interest. Second, we write a model connecting the value-added amenity to adult income. Third, we make assumptions about the representativeness of observed grades. Then we show the results, and use them to interpret the magnitudes from the previous correlational exercise. Finally, we use this to assess possible explanations of the findings in Section 4.

Definition of key parameters Let $W_c(P)$ be the average effect of growing up in county c on income at age 26 for household type P. Chetty and Hendren estimate $var(W_c(0))$ and $var(W_c(1))$, with P defined such that $P = 1$ for households above the national median level of earnings and $P = 0$ otherwise.[26]

Let $W_c^{TS}(P)$ be the effect only through the value-added amenity (i.e. through a county's test-producing inputs) of growing up in county c on income at age 26 for a child from household type P. We would like to learn $var(W_c^{TS}(0))$ and $var(W_c^{TS}(1))$

[25]This may be less persuasive if the factors which determine test score effects change more quickly than the factors which determine test score levels.

[26]We are only interested in the *variance* of place effects, so the effect $W_c(P)$ can be defined relative to any counterfactual so long as the counterfactual is the same for all counties.

under assumptions about the effects on incomes of the inputs which produce test scores.

Model of earnings effects We assume that the effect on income of the value-added amenity is equal to the quantity of the amenity provided across all K-12 grades, multiplied by a rate of return to that amenity. That is,

$$W_c^{TS}(P) = R\sum_{g=0}^{12} TS_{cg}(P),$$

where R is a rate of return to the value-added amenity and $TS_{cg}(P)$ is the average quantity of the value-added amenity provided to students from household type P in county c and grade g. Our measure $countyVA_{cg}^{P}$ from Section 3 is an estimate of $TS_{cg}(P)$.

This assumes that the return to the value-added amenity is the same across grades, consistent with Chetty et al. (2014b) if differences in test score production were due to teacher quality. We also assume that R is not a function of household type, since Chetty et al. find similar absolute returns across parental income levels, and we assume that R is identical across all counties. Finally, we are assuming that there are not diminishing marginal returns to the value-added amenity. Once again, this is consistent with Chetty et al.[27]

We are interested in $var(W_c^{TS}(P))$. Using the previous equation, we can write

$$\begin{aligned} var(W_c^{TS}(P)) &= R^2\ var(\sum_{g=0}^{12} TS_{cg}(P)) \\ &= R^2\left[\sum_{g=0}^{12} var(TS_{cg}(P)) + 2\sum_{g=0}^{11}\sum_{g'=g+1}^{12} cov(TS_{cg}(P), TS_{cg'}(P))\right]. \end{aligned}$$

Because we do not observe students in all grades, we cannot directly estimate TS_{cg} in unobserved grades using the $countyVA$ measure. However, we can still estimate $var(W_c^{TS}(P))$ under assumptions about the representativeness of the observed grades.

Representativeness assumptions We make two assumptions about the representativeness of the observed grades. First, we assume that the average variance of $TS_{cg}(P)$ in unobserved grades g is equal to the average variance in observed grades. That is, our first representativeness assumption is as follows:

[27] Counties may have heterogeneous effects on test scores across different types of students. If the structural equation is linear at the individual level with the same rate of return for all individuals, or with the rate of return uncorrelated with an individual's test score effect, then our model equation, which relates county parameters rather than individual parameters, will hold as well.

Assumption 1. *(Representative Variances)* For each value p of P,

$$\frac{1}{13}\sum_{g=0}^{12} var(TS_{cg}(p)) = \frac{1}{5}\sum_{g=4}^{8} var(TS_{cg}(p)).$$

Second, we assume that the average covariance of the test score amenity in pairs of grades in which we do not observe at least one grade is equal to the average covariance of the test score amenity in pairs of grades where both grades are observed. That is,

Assumption 2. *(Representative Covariances)* For each value p of P,

$$\frac{1}{78}\sum_{g=0}^{11}\sum_{g'=g+1}^{12} cov(TS_{cg}(p), TS_{cg'}(p)) = \frac{1}{10}\sum_{g=4}^{7}\sum_{g'=g+1}^{8} cov(TS_{cg}(p), TS_{cg'}(p)).$$

Inputting these assumptions to the equation for $var(W_c^{TS}(P))$ gives

$$var(W_c^{TS}(P)) = R^2\left[\frac{13}{5}\sum_{g=4}^{8} var(TS_{cg}(P)) + \frac{78}{5}\sum_{g=4}^{7}\sum_{g'=g+1}^{8} cov(TS_{cg}(P), TS_{cg'}(P))\right].$$

The assumption that observable covariances are representative is likely the stronger of these two assumptions. Using the *countyVA* measure, we find that variances of test scores effects are similar across observed grades, but covariances are slightly larger for closer pairs of grades.

Value of R The value of R is not known. We will argue by contradiction that R is small, and in particular it is smaller than would be expected if teacher quality accounted for counties' test score effects. To construct this argument, we begin by assuming R is *not* small.

We do this by quoting estimated returns to high value-added teachers from Chetty et al. (2014b), who estimate that a teacher who increases students' test scores by .1 standard deviation will increase students' incomes in early adulthood by approximately 1%.[28]

Chetty et al. report that their estimate is derived from a population in which a standard deviation of test achievement is approximately equal to a standard deviation of test score achievement nationwide, as measured through NAEP scores. Since this is true of North Carolina as well, there is a natural mapping between standard deviations of student achievement in Chetty et al. and in our context.

[28]More precisely, Chetty et al. report that a one standard deviation higher value-added teacher increases test achievement by .13 standard deviations, and also increases earnings in early adulthood by 1.34%, so we set R to be 1.34% higher income for each .13 standard deviations that test score achievement is increased.

An alternate measurement might have been the returns to classroom quality as measured by Chetty et al. (2011), who find that a one standard deviation better kindergarten classroom at increasing test scores leads to an increase of close to 3% in adult earnings. However, their data comes from a disadvantaged population in Tennessee, and we do not know of a clear mapping between standard deviations of student achievement on the exams used by Chetty et al. (2011) and achievement on the exams used for the North Carolina data. Nonetheless, unless standard deviations of classroom effects were much larger in Tennessee than elsewhere, it is likely that use of these results might have yielded a larger estimate for our variance of interest, $var(W_c^{TS}(P))$. Correspondingly, our ultimate conclusion that the returns to test scores when produced by counties are lower than the returns when produced by teachers would be strengthened by using this estimate instead.

Note that these existing estimates of R are estimates of the return to an input which produced a test score change, not of the return to test score production per se, since teachers and classrooms which are good at producing test scores may be unusual in other ways. Jackson (2017) finds that teachers who are good at producing test scores are also on average good at producing non-cognitive outcomes. Chetty et al. (2014b) find that high value-added teachers have effects on long-run non-cognitive outcomes but little impact on long-run test scores, while Chetty et al. (2011) find fadeout of test score effects but less so of non-cognitive effects of high value-added classrooms.

Model results We estimate the variances and covariances of $TS_{cg}(p)$ used in Assumptions 1 and 2 using the sample variances and covariances of $countyVA_{cg}^{p}$.[29] Table 5 shows the resulting estimates of the standard deviation of $\sum_{g=0}^{12} TS_{cg}(P)$. That is, it is the standard deviation of the sum across grades of county effects on test scores. Results are presented for P being either eligible for free/reduced price lunch (first column) or not eligible (second column), and with test scores measured as either math, reading, or total achievement (normalized sum of normalized math and reading scores). The units are standard deviations of single-year student performance. For comparison, a one standard deviation above average teacher increases student test scores by a little over .1 standard deviations. So, the difference between an average county and a one standard deviation above average county is approximately like having a one standard deviation better teacher in four different grades.

Table 6 multiplies the estimates for total scores by the estimates of Chetty et al. (2014b) to produce an estimate of the standard deviation of $W_c^{TS}(P)$, the effect on income of living in county c that arises due to improvements in test score performance.

According to these estimates, the effect of living in a one standard deviation

[29] $countyVA$ is estimated and therefore may vary to some extent simply due to sampling error. However, there are many observations within each county, such that sampling error accounts for only a very small fraction of variation in $countyVA$.

Table 5: Standard deviation of $\sum_{g=0}^{12} TS_{cg}(P)$

	Free/reduced	Not free/reduced
Total	0.388	0.483
	(0.036)	(0.045)
Math	0.454	0.611
	(0.047)	(0.046)
Read	0.441	0.515
	(0.034)	(0.058)

Bootstrapped standard errors in parentheses.

better county for the value-added amenity is to increase earnings in early adulthood by 4% for children from low-income families, and by 5% for children from high-income families. This can be compared with Chetty and Hendren's findings that, nationally, a one standard deviation better county in terms of effects on income raises incomes by 10% for low-income and 6% for high-income children.

Within North Carolina, Chetty and Hendren's estimates suggest that a one standard deviation better county in terms of effects on income raises incomes by 6% for low-income and 4% for high-income children. (We obtain this number by multiplying the national 10% and 6% numbers reported by Chetty and Hendren by the ratio of the standard deviations of county effects on household income percentile for counties in North Carolina to the same standard deviation for counties nationwide.) In other words, the model's estimate of $var(W^{TS})$ seems close to Chetty and Hendren's estimates of $var(W)$ – that is, suggesting that test scores alone should be able to explain virtually all geographic variation in effects of places on incomes in adulthood.

Table 6: Standard deviation of $W_c^{TS}(P)$

	Free/reduced	Not free/reduced
Total	4.00%	4.98%
	(0.371)	(0.464)

Bootstrapped standard errors in parentheses.

The variance of test score effects is surprisingly large even without considering the lack of positive correlations found in Section 4. Prior work suggests that a significant majority of the variation in quality even of *school system* inputs is not captured by

effects on test scores (Chetty et al. 2011, Chamberlain 2013, Jackson 2017),[30] and one might expect community quality to be less tied to test score production than school system quality.

Interpretation of coefficients We can also use the model to interpret the coefficients in Table 3.

Suppose that R is as above, i.e. that a .1 increase in $Test_c^P$ should correspond to a 1% effect on income. If there were no correlation between test score production in grades 4 through 8 and in other grades, we would expect the coefficients in Table 3 to be .16 for low-income kids (Columns 1 and 2) and .28 for high-income kids (Columns 3 and 4). Under the assumption of representative covariances used in our model (Assumption 2), these numbers become .26 for low-income kids and .51 for high-income kids.

These numbers are calculated as follows. Suppose a .1 increase in $Test_c^P$ corresponds to a 1% effect on income. Chetty and Hendren report that, at the 25th percentile of parental income, an exposure effect of .16 percentile points can be roughly translated to a .5% change in income. At the 75th percentile, they report that a .17 percentile point effect can be translated to a .3% change in income. Furthermore, the left-hand side variable in the regression of interest is the total causal effect of living in a particular county divided by 20 years of exposure. Using these conversions, the expected magnitude of the coefficients in Panel A of Table 3 should therefore be approximately .16 for low-income kids and .28 for high-income.[31]

We then further adjust this estimate for the expected correlation between test score effects in unobserved grades and the effect in observed grades. Fitting a linear model, the expected increase in test score effect in grade g' not between 4 and 8, $TS_{cg'}(P)$, given a single-unit increase in the sum in observed grades, which we estimate with $Test_c^P$, is

$$\frac{cov(\sum_{g=4}^{8} TS_{cg}(P), TS_{cg'}(P))}{var(\sum_{g=4}^{8} TS_{cg}(P))}.$$

The covariance term is equal to $\sum_{g=4}^{8} cov(TS_{cg}(P), TS_{cg'}(P))$. Applying Assumption 2 allows us to estimate this value. Combined with our estimate of the variance in the denominator, and multiplying by 8 to reflect that 8 of the 13 grades between 0 and 12 are omitted, the total test score effect in unobserved grades for $P = 0$ is estimated to be .59 units higher for each unit increase in $Test_c^0$, and .80 units higher

[30] Furthermore, even inputs which increase test scores may not be affecting long-run outcomes through test score production itself, especially since test score gains tend to fade out (e.g. Heckman 2008, Deming 2009, Chetty et al. 2011, Chetty et al. 2014b).

[31] Suppose a .1 increase in $Test$ leads to a 1% increase in income. Then a 1 unit increase in $Test$ is a 10% increase in income, which is a .5% increase in income per year of exposure, which is a .16 percentile point gain in income per year of exposure. Similarly, the number for the 75th percentile of parental income is $10\% * \frac{1}{20} * \frac{.17}{.3\%} = .28$.

for $P = 1$ for each unit increase in $Test_c^1$. Multiplying the numbers obtained without the representative covariances assumption by 1.59 and 1.80 yields the implied coefficients of .26 for low-income kids and .51 for high-income kids.

For low-income kids, the actual coefficient estimates (Columns 1 and 2 of Table 3) are around .05, which is lower than the expected coefficient of .26. However, the expected coefficient cannot be statistically rejected.

For high-income kids, the coefficient estimates (Columns 3 and 4) are around $-.24$. That is, the measured coefficient has approximately half the magnitude of the anticipated coefficient, with the opposite sign.

Interpreting correlational results The purpose of the model was to help understand the lack of positive correlation between counties' effects on test scores and on incomes. We previously described three possible explanations: (i) that counties do not differ substantially in their production of test scores, (ii) that there is omitted variables bias, and (iii) that the inputs which produce differences in test scores are not productive for income.

The results suggest that, at the rate of return from the literature on teacher quality, variation in the value-added amenity alone can account for about the same amount of variation in place effects on income as is measured by Chetty and Hendren. In other words, differences in test score production across counties are large.

Even large differences are not necessarily statistically detectable, however. For low-income kids, the lack of correlation could plausibly be due simply to large standard errors. For high-income kids, on the other hand, the lack of positive correlation is not simply because of sampling error, as the main regression results reject the expected coefficient from a distance of nearly 10 standard errors.

In other words, the first explanation – that there is insufficient variation in test score production to be detected in our sample – is plausible for low-income children, but not for high-income children.

We can also consider what the model implies about the plausibility of the second explanation above, omitted variables bias. Suppose the true structural equation for $W_c(P)$ is

$$W_c(P) = W_c^{TS}(P) + U_c^P,$$

where U captures the influence of all amenities other than the amenity which produces test scores. If U were correlated with test score production, we would have omitted variables bias as defined in Section 4. How severe would this omitted variables bias need to be to produce the patterns observed in the data?

As a stylized version of the results from Section 4, suppose that income production is uncorrelated with test score production, and therefore $W_c(P)$ is uncorrelated with $W_c^{TS}(P)$. As a stylized version of our model results, suppose that $var(W_c(P)) = var(W_c^{TS}(P))$. To produce these two facts in tandem, the correlation between U and W^{TS} would need to be $-.71$. (See Appendix C for a proof.)

There are two reasons to believe that such a correlation is implausible. First, as described in Section 4, community characteristics seem hardly correlated at all with test score production, let alone so strongly correlated, and there are ex ante reasons to have anticipated a *positive* correlation.

Second, and more fundamentally, such a strong correlation would suggest that U does not really fit our definition of an omitted variable. For our purposes, an omitted variable is something which is not directly involved in the production of test scores. For example, the quality of local politicians might be an omitted variable (since it is only indirectly related to test score production), while teacher quality would not be. It is difficult to imagine that U could be so strongly correlated with test score production without actually being directly involved in test score production. For comparison, the correlation between our estimates of production of test scores for different income types ($Test_c^0$ and $Test_c^1$) is .76.

Of course, the rate of return R may not be exactly the value used in the model above. (Or, the value of R might be effectively lower given the possibility of attenuation bias discussed in Section 4.) Suppose that the true value of R were $1/k$ times as large as the value we use, for some k. Keep the stylized description that there is no correlation between $W_c(P)$ and $W_c^{TS}(P)$. Then the required correlation between U and W^{TS} would be

$$\frac{-1}{\sqrt{k^2+1}}.$$

See Appendix C for the proof.

This means, for example, that if the true R were half as large as the value we quote from Chetty et al. (2014b), the correlation between test score production and omitted income-producing variables U would need to be $-.45$ – still a very strong negative correlation. At one-third of the value of R we quote, the required correlation would be $-.32$. To push the required correlation past, say, $-.2$, R would have to be no more than one-fifth of the quoted value.

This suggests that omitted variables bias alone is not a suitable explanation for the results from Section 4. Even if test score production is substantially negatively correlated with unobserved determinants of income in adulthood (which we have no basis to believe, either theoretically or based on the correlation of test score production with observed determinants of income in adulthood), R is likely to be lower than would be expected if counties' test score production differed due to teacher quality.

This conclusion is much stronger for high-income children, for whom this statement of stylized facts is actually conservative, than it is for low-income children. For low-income children, as described above, a sampling error argument suffices – though the low point estimate of the correlation of interest and the results for high-income children both give pause.

This exercise is robust to several concerns, at least for high-income children. Attenuation bias might bias the correlation between counties' effects on incomes and test scores in the direction of zero – but for high-income children, bias in the direc-

tion of zero means the estimated correlation would be *higher* (less negative) than the true correlation. The representative covariances assumption used in constructing the model might lead to an overestimate of the variance of W_c^{TS} – but without it, the expected coefficient in Table 3 has roughly the same magnitude as the measured coefficient, with the opposite sign. Incorrect identifying assumptions by Chetty and Hendren may bias their results – but this would most likely mean they have overestimated the variance of $W_c(P)$, which in turn implies that the variance of $W_c^{TS}(P)$ is larger by comparison, which thus requires an even stronger correlation of $W_c^{TS}(P)$ with U to explain the results from Section 4.[32]

Summing up, the model suggests that, at least for high-income kids, the correlational results described in Section 4 are not plausibly due to omitted variables bias or lack of variation in test score production. We conclude from this that the rate of return R must be lower – probably substantially lower – than the value we would expect if differences in test score production were due to teacher quality.

6 Interpretation as effects of schools

Our research question is whether counties' test score production is an indication of whether schools there are preparing students for the workforce. So far, we have presented evidence that this is not the case because counties' test score production seems not to be particularly helpful to students' incomes in adulthood.

Next, we consider which, if any, school inputs could even play a role in determining counties' effects on test scores. It is widely believed that teachers are the most important input provided by the school system. Yet the conclusion of the previous section – that a certain amount of test score production by counties does not increase adult incomes by as much as the same amount of test score production generated by teacher quality – would not make sense if counties differ in their test score production only due to differences in teacher quality.

Therefore, we begin by assessing whether it is plausible that differences in counties' production of test scores are due to amenities other than teacher quality.

Teacher switching across counties We assess the importance of non-teacher amenities to counties' test score production by observing teachers who move between counties.

The logic behind looking at teacher moves between counties is as follows. Suppose that county A has a non-teacher amenity which results in test scores θ higher than in county B. Then we would expect that a teacher in county A would have an estimated value-added θ higher than the same teacher would have had in county B. If we assume that teachers' actual quality is stable across years – or that year-to-year changes in

[32] An exception to this possibility would be if Chetty and Hendren's results underestimate the variance of place effects because of their reliance on switchers, who may sort into similar amenities in all locations, as we document in Appendix A.

quality are uncorrelated with the direction of teachers' moves – then we can measure θ by observing the average change in the same teachers' measured value-added as they move between the two counties.

We implement this design as follows. Suppose teacher i taught at county c' in year $t-1$, but switched to county c in year t. We estimate a specification that regresses the differences in the value-added of a teacher who switches counties, on the differences in the value-added of the old and new counties, as follows:

$$\Delta teacherVA_{it} = \alpha_0 + \alpha_1 \Delta countyVA_{it} + u_{it},$$

where, letting $teacherVA_{it}$ be the sample average of the residual $studentVA$ across all students that teacher i teaches in year t, and letting $countyVA_{ct}$ be the average student residual in the entire county c in during year t,

$$\Delta teacherVA_{it} := teacherVA_{it} - teacherVA_{it-1}$$

and

$$\Delta countyVA_{it} := \sum_{\tau=2000}^{t-1} countyVA_{c\tau} - \sum_{\tau=t}^{2006} countyVA_{c'\tau}.$$

Note that this captures the change in quality between counties using only years before the teacher entered to measure value-added in the new county, and only years after the teacher left to measure value-added in the old county. Measuring county value-added in this way, such that the particular teacher is excluded from the measure, eliminates any correlation between teacher and county residuals due to county-year shocks, and therefore isolates only the influence of long-term amenities which differ across counties.[33]

We restrict this analysis to elementary school grades, since students in these grades are generally taught by a single teacher. Matching of students to teachers in the data is imperfect, so we perform additional robustness checks, described in Appendix D, to verify that our results in this section are not driven by erroneous assignment of teachers to students in the data.

If differences in teacher quality completely explain variation in value-added across counties, then there is no scope for non-teacher amenities to vary across counties, and we would see that α_1 in the equation above would be equal to 0. On the other hand, if differences between counties in non-teacher amenities completely explained the variation in the value-added amenity, then we would expect to see teachers' estimated value-added would jump by the difference in county value-added at the time that a teacher switches counties, and therefore α_1 would be equal to 1.

The first two columns of Table 7 show our results from this regression, run separately for math and reading value-added measures. It appears that differences in counties are not primarily driven by differences in teacher quality. For both math

[33]This also excludes the teachers themselves from the measurement of $countyVA$, and mitigates any concern that teachers affect the performance of their peers.

and reading, we can rule out that α_1 is equal to zero at the .01 significance level. These results suggest that there are amenities other than teacher quality which affect test scores, which differ across counties, and which are stable across time. Furthermore, such amenities likely account for the majority of the differences in county effects on test scores. For example, among county pairs which differ by 1 unit of the value-added amenity, our results suggest that non-teacher amenities account for around .8 units of that difference – perhaps a bit more for reading, or less for math. That is, the large majority of county effects on test scores are due to differences in non-teacher amenities.

Table 7: Change in teacher value-added when moving

	Switch counties		Switch schools	
	$\Delta teachVA_M$	$\Delta teachVA_R$	$\Delta teachVA_M$	$\Delta teachVA_R$
Δ *Math*	0.693**		−0.104	
	(0.140)		(0.055)	
Δ *Reading*		0.937**		−0.059
		(0.145)		(0.077)
N	584	583	832	829

** indicates $p < 0.01$. Robust standard errors in parentheses. Regressions are estimated for teachers in grades 4 and 5 with at least 10 and no more than 40 students per class. The first two columns correspond to teachers who switch counties, and the last two columns correspond to teachers who switch schools within the same county.

One threat to this design would be if teacher quality were not stable, as suggested by Chetty et al. (2014a), and would have trended in the direction of the change in county quality even if the teacher had not moved. As a placebo test to detect such a threat based on existing trends in teacher performance, we regress changes in the teacher's value added in years before and after they move on the difference in the value added between the old and the new county (measured, as in the previous regression, using only years when the teacher is not yet or is no longer present in a county). That is, suppose a teacher moves from county c to county c'. We estimate

$$\Delta teacherVA_{it+m} = \alpha_0^m + \alpha_1^m \Delta countyVA_{ict} + u_{it}^m,$$

where $\Delta teacherVA_{it}$ and $\Delta countyVA_{ict}$ are defined as above and m is an integer.

We estimate this specification for each year starting three years before the teacher switches and up to three years after the switch, i.e. for m between -3 and 3.[34] If it is the case that teachers selectively switch across counties such that teachers who are generally experiencing an improvement in value-added over time move to better counties, then we should measure α_1^m to be positive for years before and after the

[34]Given the limited number of years of data we use, it is mechanically impossible for a teacher's change in value-added to be observed seven times. Therefore, the results come from an unbalanced panel.

move, i.e. for $m \neq 0$. Figures 1 and 2 plot our estimates of the coefficient α_1^m against m by subject, with 95% confidence intervals. For most years other than $m = 0$ (the year of the actual move), the coefficients are statistically indistinguishable from zero, but the placebo point estimates are uniformly negative. That is, the jump in value-added at the time of the move is not similar to trends in value-added in other years; if anything, teachers who move to a better (worse) county tend to have a modest downward (upward) trend in their performance, except in the year of the move. While the apparent non-randomness of moves is somewhat disconcerting, the results in the figures are only consistent with an upwards bias in Table 7 if the non-randomness produces an upwards bias exclusively in the first year of the move, i.e. if the decision to move is correlated with a shock to performance which is realized only in the period after the decision to move has been made.

This suggests that, if anything, our previous results may *overstate* the importance of teacher quality in explaining differences in county effects on test scores. In other words, very little of the variation in counties' effects on test scores is due to teacher quality.

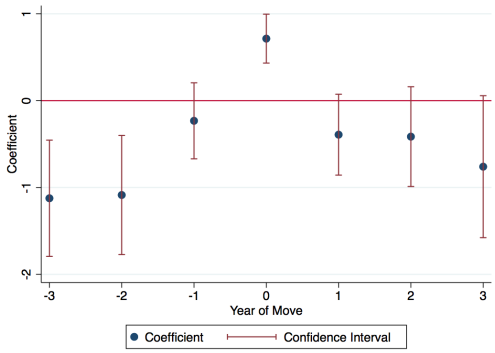

Figure 1: Math Scores

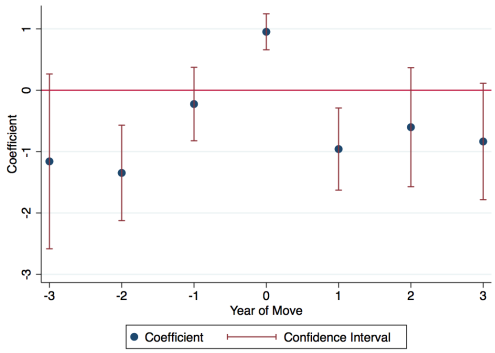

Figure 2: Reading Scores

Teacher switching across schools The jump in teacher value-added when moving counties is surprising, as previous research (Chetty et al. 2014a) has found that teachers' estimated value-added remains stable as they move across schools, which has been interpreted as evidence that value-added measures are unbiased. However, this prior research focuses on teacher moves within concentrated geographic areas, and would not necessarily detect bias due to differences in amenities which vary across counties.

Consistent with this prior research, we confirm in our data that teachers who switch across schools within the same county do not experience sudden changes in measured value-added. This demonstrates that our results from observing teachers who move across counties are not inconsistent with prior research.

Furthermore, this suggests that within-district variation in school-level characteristics is unimportant to student test scores, conditional on teacher quality. If

school-level amenities are not important for determining student test scores, then they are unlikely to explain differences in student test score production across counties either.

We implement this school-switching design as follows. Suppose teacher i taught at school s' in year $t-1$, but switched to school s within the same county in year t. We estimate

$$\Delta teacherVA_{it} = \alpha_0 + \alpha_1 \Delta schoolVA_{ist} + u_{it},$$

where, letting $teacherVA_{it}$ be the average residual across all students that teacher i teaches in year t and $schoolVA_{st}$ be the average student residual in the school s that teacher i teaches in during year t,

$$\Delta teacherVA_{it} := teacherVA_{it} - teacherVA_{it-1}$$

and

$$\Delta schoolVA_{ist} := \sum_{\tau=2000}^{t-1} schoolVA_{s\tau} - \sum_{\tau=t}^{2006} schoolVA_{s'\tau}.$$

Analogously to the county-switching quasi-experiment above, this constructs $schoolVA$ only using those years in which the teacher of interest is not present in the school.

The results are shown in the last two columns of Table 7. We cannot reject that teachers do not change in value-added as they move across schools.

Since school-level amenities such as principal quality do not seem to appreciably affect test score achievement (holding the teacher constant), it is unlikely that they can explain why value-added is higher in some counties than in others.

County boundaries Since teacher quality and other amenities which vary across schools within district do not seem to explain county differences in test score production, the question remains what, precisely, accounts for these differences. While non-school inputs such as pollution or culture might be important, we lastly consider some suggestive evidence about whether differences in the school system at a level higher than schools may be important. For example, school districts may prioritize certain kinds of instruction at all schools, may differ in the support offered to staff, or may differ in the quality of facilities that they generally provide.

To disentangle the effects of geography broadly from the effects of school district assignment specifically, we measure changes in value-added around county boundaries. In North Carolina, school districts usually coincide exactly with a county.[35] If administrative assignments to school districts are responsible for the differences in

[35] A few other school districts coincide with cities or with other portions of counties, and North Carolina's charter schools operate outside of the normal school district system. We do not exploit these additional district boundaries because schools in these districts are almost all close to district boundaries, complicating the desired interpretation of taking a limit as the border is approached.

Table 8: Change at county boundaries

	5 km		10 km		20 km	
	$School_M$	$School_R$	$School_M$	$School_R$	$School_M$	$School_R$
$County_M$	0.868**		0.788**		0.861**	
	(0.312)		(0.189)		(0.067)	
$County_R$		0.457		0.669**		0.761**
		(0.351)		(0.238)		(0.126)
N	73	73	728	728	7,909	7,909

** indicates $p < 0.01$. Robust standard errors clustered by county in parentheses. All regressions control for year and grade fixed effects.

value-added across counties, we would expect to see sharp jumps in value-added at county boundaries.

Unfortunately, the small number of switching teachers does not permit us to precisely estimate a jump in non-teacher amenities at county boundaries. However, unless there are substantial jumps in teacher quality at county boundaries – which seems unlikely, given the modest differences in teacher quality across counties – we might be willing to believe that any jump in school value-added near county boundaries gives a reasonable approximation of the influence of district-level amenities.[36]

We study jumps in value-added at county boundaries as follows. We find school pairs which are within a certain radius D of each other but in different counties, and regress the difference in the average test score residual of the students in the schools on the difference in the average test score residual of the students in the counties that the schools belong to, using the specification

$$\Delta schoolVA_{ic,jc',t} = \alpha^D \Delta countyVA_{cc't} + u_{ijt},$$

where $\Delta schoolVA_{ic,jc',t}$ is the difference in the average residuals (across all grades) of schools i in county c and j in county c' in year t, and $\Delta countyVA_{cc't}$ is the difference in the average residuals (across all grades) of counties c and c' in year t. The constant is suppressed because the ordering of schools is arbitrary. We estimate this specification multiple times for different degrees of proximity between school pairs, i.e. different values of D.

Our results are shown in Table 8. A coefficient of 1 would indicate that the entire difference in county value-added exists among school pairs within distance D of each other, while a coefficient of 0 indicates that the difference has entirely disappeared. For reading, the point estimates of differences in test score production appear to be shrinking as we consider pairs closer and closer to the county boundary. However, while the estimates are not statistically different from zero, they remain substantially positive even over close distances, and we cannot reject a large number. For math,

[36]Such a jump may also represent county-level amenities which are not implemented by the school district.

the differences are stable and large at all distances, though the confidence interval is wide at 5 kilometers. This is consistent with some role for school district (or other county) policies and management practices in determining county test score value-added. Indeed, coefficients on the order of .8 are comfortably within the confidence intervals, meaning that differences in school districts could be large enough to explain the change in teacher value-added when teachers move across counties. However, the estimates for reading are suggestive that factors other than the school district may play some role as well.

We consider this analysis to be merely exploratory, for three reasons. First, at short distances between schools, our estimates suffer from a lack of statistical power. Second, even when schools are very close to each other, children may be drawn from further away, since students generally attend schools in the same county in which they reside. Therefore, supposing test score differences were solely driven by some other characteristic of a place (e.g. culture or pollution), we might still expect to see some jump in the above regression, even with no role at all for the school system. Third, we cannot rule out that relevant inputs other than school district also change at county boundary lines.

A natural question is what other amenities might explain differences in county value-added, if not school system inputs. Some possibilities include cultural factors, environmental factors (e.g. air and water quality), or other community-wide social amenities. Another possibility is that family-level inputs change in response to the community in which the family lives; this would be picked up in our empirical analysis as an effect of the community, even though the effect occurs via family inputs. Future research may be able to provide more guidance about any non-school system sources of test score effects.

7 Discussion

The goal of this paper is to assess the common belief that a county's production of test scores is a meaningful signal about how well the school system is preparing students for the workforce. The implicit logic behind this belief is that (i) differences in counties' test score performance are probably mostly due to school system inputs, and (ii) school system inputs which increase test scores also presumably substantially increase incomes in adulthood.

Even beyond the obvious objection that different sorts of people live in different places, and therefore test score levels vary for reasons other than local amenities, we find that the case for (i) is surprisingly shaky. First, in Section 6, we find that teacher quality accounts for only a fraction of the difference in test score production across places. Second, we find that school-level inputs such as principal quality are unlikely to matter much either. This leaves only district-level factors as potentially important school system inputs driving counties' effects on test scores. Our exploratory analysis suggests that school districts may be able to account for a large fraction of differences

in test score production – perhaps even the entire part beyond teachers – but the statistical evidence is not conclusive.

Granting belief (i) only leads to the conclusion that belief (ii) is likely to be incorrect. For high-income students in particular, the evidence is strong that the inputs which account for county-level variation in test score production do not have a large positive effect on incomes in adulthood.

We conclude that claims (i) and (ii) are not jointly correct in the population we study.

We additionally conclude from the results in Section 6 that teacher value-added measures become biased when comparing teachers in different counties. For any policy regime such that comparison of personnel across regions is important, policy-makers may wish to correct estimates of teacher value-added for place differences which are detected through changes in value-added among teachers who move.

Possible mechanisms Our data do not allow us to identify precisely why the return to test score production might be low, beyond pointing to the lack of a role for teacher quality. We leave it to further work to resolve the mechanism more precisely. There are three broad possibilities: (i) that production of tested skills displaces production of non-tested skills; (ii) that short-run gains in the tested skills do not persist through adulthood; and (iii) that the tested skills are not well-rewarded in the job market.

The first possible explanation is related to common concerns about "teaching to the test." Suppose there are some human capital inputs which produce gains on test scores, while others do not. Counties' effects on test scores may reflect substitution between these inputs subject to some constraint, e.g. a finite amount of class time. If inputs are chosen roughly to maximize income in adulthood subject to the constraint on investment, then, by the fact that the resulting choice is the constrained optimum, counties which deviate by aligning curriculum or time expenditures more closely with the exam will not generate improvements in adult incomes. One reason this mechanism might be more important for counties than for teachers (justifying the discrepancy with Chetty et al. 2011 and Chetty et al. 2014b) would be if most of the variation in county value-added comes from substitution between inputs, while most of the variation in teacher value-added comes from talent, i.e. from relaxing the optimization constraint.

Certainly, our estimates from observing county-switching teachers do not point to large differences across counties in educator talent. However, in certain institutional environments, it is possible that such differences would emerge (e.g. Biasi 2018b), in which case we would expect our findings to change.

The second possible explanation is that gains in literacy and numeracy do not persist into adulthood. A common finding in the literature is that interventions which produce short-term improvements in test scores do not tend to produce large long-run gains in test scores (e.g. Deming 2009, Chetty et al. 2011, Chetty et al. 2014b). While other interventions which produce increases in short-run test scores

have also been shown to increase incomes in adulthood (e.g. Deming 2009, Chetty et al. 2011, Chetty et al. 2014b), the fadeout of test score gains combined with persistent improvements in skills not measured by standardized tests suggest that the test score gain per se might not be the important part of these interventions. One reason our results are incongruent with the literature on the return to high value-added teachers might be that counties' production of test scores is less closely correlated with production of skills not measured by standardized tests.

The final possible explanation is that literacy and numeracy are not well-rewarded in the labor market. This might occur, for instance, if they are so widespread thanks to universal education that they are effectively not scarce in the job market. In Appendix E, we document using data from O*NET and the American Community Survey that people working in occupations which require high levels of numeracy and literacy earn more than people who do not; for both skills, a one standard deviation increase in the level of skill required is associated with roughly 30% higher income. Yet simply controlling for the level of deductive reasoning required in the occupation nearly eliminates this correlation. While far from conclusive, this at least suggests it is possible that literacy and numeracy per se do not command a large premium in the labor market; instead, the wage premium of white collar jobs may be due to broader cognitive ability. If the interventions which account for counties' test score effects operate by narrowly altering literacy and numeracy without having broader effects on reasoning ability and cognitive function, such interventions would then not produce substantial gains in adult income.

Caveats There are a few important caveats to our main conclusion.

First, our results do not imply that schools do not matter, and should not be interpreted as claiming that the school system is unimportant. A body of previous research suggests that schooling produces human capital (e.g. Card 1999, Clark and Martorell 2014). Furthermore, schools produce skills which are not measured by standardized test score performance (e.g. Heckman 2008, Chetty et al. 2011, Chamberlain 2013, Jackson 2017). School inputs which produce test scores often (e.g. Deming 2009, Chetty et al. 2011, Chetty et al. 2014b, Lavy 2016) though not always (e.g. Deming et al. 2016, Dobbie and Fryer 2017) produce substantial increases in incomes in adulthood.

Second, there is no reason to assume that our results would hold in all contexts at all times. It is known that there are school system inputs, such as effective teachers, which improve both test scores and incomes. Our finding should be interpreted as a note of caution: test scores do not necessarily speak to the quality of the school system. It is important to pay attention to what inputs, exactly, account for variation in test score production.

Third, the evidence for a low return to test score production is much stronger for children from high-income families than for children from low-income families. We believe that the evidence for high-income children suggests that the return to test score production for low-income children might also be low. However, it is

possible that there are important differences either in the amenities which produce test scores for these two groups, or in the two groups' responses to the same amenity. For example, Deming et al. (2016) find that the same intervention – accountability pressure – which improved test scores was beneficial to the long-run outcomes of students at low-test score schools, but harmful to outcomes of students at high-test score schools.

Fourth, to the extent that it is desirable to generalize results, it is natural to wonder whether North Carolina is representative of the country as a whole. The primary concern would be that the inputs which account for variation in test score effects of counties within North Carolina are different from the inputs which account for variation in test score effects of counties nationwide. We cannot directly comment on this form of representativeness, since we do not have comparable individual-level panel data with a common test across many states. It is at least somewhat reassuring that, on national exams, the distribution of test score performance in North Carolina is not atypical, which would be less likely to be the case if the variation in test score production inputs in North Carolina were highly unusual (following the argument of Penney 2017).[37] However, this may occur for other reasons.

Even if North Carolina is representative of other states, a shortcoming of using data from a single state is that we cannot make any statement about the comparison of counties in different states. This matters if (i) state-level policies affect test score production, and (ii) such policies have a different return to test score production than holds for intra-state sources of variation in test score production. We leave assessing the importance and value of state-level policies to future research.

Comparison with Rothstein We now revisit our comparison with Rothstein (2018), which we began in Section 1.

Rothstein is interested in why parental income seems to be transmitted to children more strongly in some counties than in others. In the process, he also finds correlational evidence suggesting that test scores might not be important for explaining counties' effects on income. In particular, Rothstein points out that test scores of high- and low-income students do not seem to grow any more similar over the course of childhood in high-mobility counties (i.e. counties where parental in-

[37] On the National Assessment of Educational Progress (NAEP), a standardized test which is given to representative samples of students across the country, North Carolina tends to have somewhat higher than average math performance and somewhat below average reading performance; e.g. data from the NCES shows that in 2005, North Carolina had average scaled scores which were 14th among US states in 4th grade math, 32nd in 4th grade reading, 18th in 8th grade math, and 37th in 8th grade reading. See https://www.nationsreportcard.gov/ for full rankings. The dispersion of performance is also not atypical: The proportions of students at different achievement levels (below basic, basic, proficient, and advanced) on the NAEP are broadly similar to the national fractions, and we find that a standard deviation of teacher value-added corresponds to roughly the same number of standard deviations of student performance as prior studies (e.g. Chetty et al. 2014a) have found in other places, which would not be the case if the variance of student achievement were atypical in North Carolina (Penney 2017).

come is only weakly correlated with child's income, as measured by Chetty et al. 2014c) than they do in low-mobility counties. Rothstein suggests that effects on test scores might therefore not be important for explaining geographic variation in intergenerational income correlations, and hypothesizes that other variables like family structure are more important.

While the motivating questions are different – Rothstein is interested in understanding the origins of intergenerational income correlations, while we are interested in evaluating a particular way of measuring school system performance – Rothstein's and our findings clearly both support the narrative that regional variation in test score production within the United States is not a useful tool for understanding regional variation in production of adult incomes.

Our findings complement Rothstein's in part by strengthening his conclusions. First, the nature of our data allows us to more explicitly argue that we have measured causal effects of places on test scores. Second, Rothstein takes care to observe that his results are correlational, since he does not observe all other amenities in a community, but our approach to the problem allows us to argue more forcefully for a causal interpretation by arguing that omitted variables bias cannot plausibly drive the results for high-income children. Finally, Rothstein's results could be rationalized by a model in which test score production is a common amenity for children of all income levels and is beneficial to everyone, but is particularly beneficial for children from low-income families. In this case, test score production would still be important for understanding intergenerational income mobility. Our results are consistent with test score production being more valuable (or perhaps, less harmful) for low-income children, but suggest that this story is unlikely to explain Rothstein's results because it would imply greater production of incomes in communities with greater test score production.[38]

We also complement Rothstein's findings by exploring the mechanism behind his findings. His findings could be generated either by a lack of variation in test score production or by a low return to test score production. Our results point to the latter explanation. Furthermore, we are able to shed light on why there is a low return by pointing out that teacher quality is not the input producing variation in test score production.

In turn, Rothstein's findings reinforce ours. Most importantly, they suggest that our results are not simply driven by using data which are not nationally representative. Second, they help confirm that our results are not driven by the years of data used. Finally, they suggest that our results are not driven by the particular standardized test being used.

[38]In principle, there is a distinction between Rothstein's outcome – a measure of correlation between parents' and children's incomes, drawn from Chetty et al. (2014c) – and our outcome, which is place effects on incomes, drawn from Chetty et al. (2017b). But Chetty et al. (2014c) report that high-mobility communities differ little from low-mobility communities in the incomes of children from high-income families, such that high-mobility communities are apparently effective at producing incomes.

8 Conclusion

Using data from North Carolina, we find evidence suggesting that test score production is not always a useful measure of how well a community's schools are preparing children for the workforce. We first show that counties which are good at producing test scores are not any better (and, for high-income children, are significantly worse) at producing incomes. Using a simple model, we argue that this is likely in large part because the inputs which are directly responsible for differences in test score production are not useful for producing incomes. We find evidence suggesting that differences in counties' production of test scores are only fractionally due to differences in teacher quality or school-level inputs. Instead, variation in test score production mostly reflects school district-level inputs or non-school system inputs. We conclude that it is unlikely that differences in test score levels across counties are due to differences in productive inputs from schools.

References

[1] Card, D. (1999). "The Causal Effect of Education on Earnings," *Handbook of Labor Economics*, 1801-1863.

[2] Card, D. and A.B. Krueger (1992). "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy*, 100 (1): 1-40.

[3] Biasi, B. (2018a). "School Finance Equalization and Intergenerational Mobility," working paper.

[4] Biasi, B. (2018b). "The Labor Market for Teachers under Different Pay Schemes," working paper.

[5] Chamberlain, G. (2013). "Predictive effects of teachers and schools on test scores, college attendance, and earnings". *Proceedings of the National Academy of Sciences*. 110 (43): 17176-17182.

[6] Chetty, R., N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR". *Quarterly Journal of Economics* 126 (4): 1593-1660.

[7] Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates". *American Economic Review* 104 (9): 2593-2632.

[8] Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood". *American Economic Review* 104 (9) 2633-2679.

[9] Chetty, R. and N. Hendren (2018). "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects". *Quarterly Journal of Economics*, 133(3), 1107-1162.

[10] Chetty, R. and N. Hendren (2018). "The Impact of Neighborhoods on Intergenerational Mobility II: County-Level Estimates". *Quarterly Journal of Economics*, 133(3), 1163-1228.

[11] Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). "Where is the Land of Opportunity: The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics* 129 (4): 1553-1623.

[12] Clark, D. and P. Martorell (2014). "The Signaling Value of a High School Diploma." *Journal of Political Economy*, 122(2), 282-318.

[13] Deming, D. (2009). "Early Childhood Intervention and Life-Cycle skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics*, 1(3), 111-134.

[14] Deming, D. J., Cohodes, S., Jennings, J., and C. Jencks (2016). "School Accountability, Postsecondary Attainment, and Earnings". *Review of Economics and Statistics* 98 (5): 848-862.

[15] Dobbie, W. and R. G. Fryer (2017). "Charter Schools and Labor Market Outcomes," working paper.

[16] Hanushek, E. and L. Woessmann (2012). "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation". *Journal of Economic Growth* 17: 267-321.

[17] Heckman, J.J. (2008). "Schools, Skills, and Synapses." *Economic Inquiry*, 46(3), 289-324.

[18] Imberman, S.A. and M.F. Lovenheim (2016). "Does the Market Value Value-Added? Evidence from Housing Prices After a Public Release of School and Teacher Value-Added." *Journal of Urban Economics* 91(C), 104-121.

[19] Jackson, K. (2017). "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes". *Journal of Political Economy.*

[20] Kane, T. J. and D. O. Staiger (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation". *NBER Working Paper 14607.*

[21] Kane, T. J., D. F. McCaffrey, T. Miller and D. O. Staiger (2013). "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment". *(Seattle, WA: Bill & Melinda Gates Foundation).*

[22] Lavy, V. (2016). "Teachers' Pay for Performance in the Long-Run: The Dynamic Pattern of Treatment Effects on Students' Educational and Labor Market Outcomes in Adulthood," working paper.

[23] Lazear, E.P. (2003). "Teacher Incentives". *Swedish Economic Policy Review* 10, 179-214.

[24] Murnane, R.J., Willett, J.B., Duhaldeborde, Y., and J.H. Tyler (2000). "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?". *Journal of Policy Analysis and Management* 19(4), 547-568.

[25] Penney, J. (2017). "A Self-Reference Problem in Test Score Normalization". *Economics of Education Review* 61, 79-84.

[26] Reardon, S.F. (2018). "Educational Opportunity in Early and Middle Childhood: Variation by Place and Age," working paper.

[27] Rothstein, J.M. (2006). "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions." *American Economic Review* 96 (4): 1333-1350.

[28] Rothstein, J. (2017). "Measuring the Impacts of Teachers: Comment". *American Economic Review* 107 (6): 1656-84.

[29] Rothstein, J. (2018). "Inequality of Educational Opportunity? Schools as Mediators of the Intergenerational Transmission of Income". *Journal of Labor Economics*, forthcoming.

[30] Schoellman, T. (2012). "Education Quality and Development Accounting". *Review of Economic Studies* 79 (1): 388-417.

[31] Wooldridge, J. (2009). "Introductory Econometrics: A Modern Approach". Nelson Education.

Appendix

A Sorting into similar quality classrooms

We use the econometric model described in Section 3 to measure the effects of county of residence on test score achievement, and test this model by observing whether students who move between counties show the expected jump in test score achievement. It turns out that, if one fails to account for sorting to similar quality of teachers and schools, the jump in test score achievement is slightly smaller than the jump implied purely by average county effects on test scores. In this appendix, we show that students who move between counties experience changes in school and teacher

quality which are smaller than the difference in average county effects on test scores would imply.

To test for the extent of this sorting, we run the following regression:

$$\Delta schoolVA_{it} = \alpha_0 + \alpha_1 \Delta countyVA_{it} + u_{it},$$

where, letting $schoolVA_{it}$ be the average student residual (excluding i) in the school that student i attends in year t and $countyVA_{it}$ be the average student residual (excluding i) in the entire county that i lives in during year t,

$$\Delta schoolVA_{it} := schoolVA_{it} - schoolVA_{it-1}$$

and

$$\Delta countyVA_{it} := countyVA_{it} - countyVA_{it-1}.$$

If families choose a school at random in each county (weighted by the number of students attending that school), then $\alpha_1 = 1$. If, on the other hand, families were able to sort into schools with identical measured value-added no matter which county they live in, then $\alpha_1 = 0$.

We also estimate an identical specification of classroom effects on county effects. Our results are shown in Table 9A and Table 9B. The columns labeled with free/reduced price lunch status use only switchers of that status, and likewise estimate county, school, and class value-added using sample averages of peers of that same status. The estimates suggest that the extent of sorting is modest but non-zero; the change in school quality faced by students who move between counties is very nearly as large as the change in average school quality between the two counties, but the change in classroom quality is somewhat smaller.

Our results in Section 6 suggest that the same teacher will generally have a very different measured value-added in different counties. So, while these estimates make it appear that the extent of sorting is quite modest, this might actually be consistent with quite substantial sorting. For example, if students sorted to literally identical teachers no matter which county they lived in, we would expect to see a coefficient around .8 in these regressions.

Table 9A: School Effects on County Effects

	All		Free/reduced		Not free/reduced	
	$School_M$	$School_R$	$School_M$	$School_R$	$School_M$	$School_R$
$County_M$	0.930**		0.939**		0.919**	
	(0.008)		(0.010)		(0.011)	
$County_R$		0.891**		0.924**		0.852**
		(0.007)		(0.011)		(0.011)
N	60,900	60,369	31,778	31,442	29,122	28,927

** indicates $p < 0.01$. Robust standard errors in parentheses.

Table 9B: Classroom Effects on County Effects

	All		Free/reduced		Not free/reduced	
	$Class_M$	$Class_R$	$Class_M$	$Class_R$	$Class_M$	$Class_R$
$County_M$	0.925**		0.950**		0.890**	
	(0.023)		(0.030)		(0.034)	
$County_R$		0.848**		0.904**		0.771**
		(0.026)		(0.035)		(0.037)
N	16,855	16,663	9,370	9,241	7,485	7,422

** indicates $p < 0.01$. Robust standard errors in parentheses. Regressions are for students moving in grade 5.

We interpret these estimates as a sign that students are sorting into similar quality classrooms no matter which county they live in. The smaller coefficient for high-income students suggests that they are able to sort more effectively, perhaps consistent with their families' greater ability to influence their child's choice of school and classroom.

Another possible interpretation of this finding, combined with the results in Table 2, is that there exist unobservables which bias the value-added equation, and that students sort to peers who are similar on those unobservables. That is, students may have unobserved characteristics which differ across counties, but do not differ between the old and new classrooms for students who move, such that the degree of bias in the value-added measure is the same in the old and new classrooms (hence the results in Table 2) but not in the old and new counties. However, if this were the case, then substantial differences in student unobservables would exist *within* counties, which would conflict with the observation that teacher value-added is stable as teachers move across schools with different measured value-added within a county (see Table 7 in Section 6). Broadly, a model in which students sort to identical unobservable quality of peers as they move between counties implies a remarkable ability to sort on the basis of those unobserved characteristics, which would seem to conflict with the existing literature on teacher value-added models (Kane and Staiger 2008, Kane et al. 2013, and Chetty et al. 2014a).[39] Regardless, some modest bias due to uncontrolled unobservables would presumably make our estimates in Section 4 conservative, since Table 4 shows much higher coefficients relating counties' test score levels and income production.

Attenuation of amenity differences due to sorting is relevant for the interpretation of Chetty and Hendren's estimates of county effects on incomes, since, under their

[39]Chetty et al. (2014a) in particular document that markers of parental income are orthogonal to estimated teacher value-added, conditional on value-added model controls. If our results were driven by sorting to peers with similar unobserved characteristics, the smaller coefficients for high-income families in Tables 9A and 9B would suggest that high-income families sort more strongly than low-income families on the basis of these unobservables, which would contradict Chetty et al.'s finding.

identifying assumptions, their estimates are valid estimates of place effects for a population of families who switch counties. In terms of comparability of our estimates to theirs, we measure the difference in average test score value-added across counties, not the average difference in value-added faced by children as they move across counties. To the extent that sorting occurs, this will increase the variance of the object that we measure relative to the variance of the object they measure.

B Placebo tests for student switchers

In Section 3, we compare the change in residuals that our value-added model suggests students should experience when they move to the changes they *actually* experience, and argue that the former is likely to be a reasonable estimate because it aligns with the latter. But if each of these exercises suffers from an equal degree of bias, then these two measures may coincidentally align without our primary measure being accurate. Therefore, in this appendix, we investigate whether the student-switching quasi-experiment is likely to suffer from bias.

In order for the student-switching quasi-experiment to yield valid estimates, it must be that the changes in residuals that students would have experienced at time t, regardless of where they lived, are not correlated with the decision of where to move. This might not true, since people are not assigned where to live by lottery.[40] In order to assess this possibility, we investigate whether students' moves seem to be non-random with respect to their change in residuals before or after the move.

We believe that, in particular, this placebo test in the period before the move is the most informative, for the following reason. Endogeneity in the student-switching quasi-experiment could arise due to non-randomness in the decision to move. But, for most families, the *decision* to move is made before the *actual* move – typically during the prior school year. Therefore, any such non-randomness of the move with respect to students' residuals would be expected to manifest itself at this time.

We implement our placebo tests as follows. In Section 3, our left-hand side variable was $\Delta studentVA_t$, the difference in a student's residual between the year they moved, t, and the year before, $t-1$. We would like to know whether students' moves are random with respect to this variable. (While student moves are surely not random, the only possibility for bias in our main regression is non-randomness with respect to the left-hand side variable.) While we cannot directly test for this non-randomness, we can at least test for randomness with respect to the comparable outcome variable the year before or after, $\Delta studentVA_{it-1}$ and $\Delta studentVA_{it+1}$ – the former of which, as described above, we consider especially informative. To test

[40] On the other hand, it is also not obvious that trends in residuals *must* be correlated with the moving decision, given the evidence that families engage in little sorting on the basis of schools' value-added – see e.g. Rothstein (2006) or Imberman and Lovenheim (2016) – and given the evidence that the controls in the value-added equation already successfully account for a rich set of family characteristics (Chetty et al. 2014a).

for non-randomness, we can regress these variables on $\Delta countyVA_{it}$. If counties which our econometric model labels as higher value-added receive movers who have significantly different trends in their residuals from counties which our econometric model labels as lower value-added, then we would expect the coefficient to be non-zero.

One complication of this placebo test is that $\Delta countyVA_{it}$ is positively correlated with county value-added in the years after the student moves, and negatively correlated with county value-added in the years before the student moves. To the extent that $countyVA$ is correlated across grades, correctness of the identifying assumptions we intend to test therefore does not necessarily imply zero correlation between $\Delta countyVA_{it}$ and $\Delta studentVA_{it-1}$ or $\Delta studentVA_{it+1}$. To correct for this, we construct the variable $diffVA_{it}$ to be equal to the difference between i's $studentVA$ in period t and the sample average of i's same-grade classmates' $studentVA$ in period t, and substitute the year-to-year change in $diffVA$ for the change in $studentVA$ on the left-hand side of our placebo regressions. Because county and average classroom value-added track each other closely (though not exactly; see Appendix A) among movers, results are not meaningfully affected by comparing students' residuals to the average of their same-grade countymates' instead of classmates'. This adjustment allows for sharp delineation between random and non-random moving: correctness of our assumptions predicts an exact zero coefficient when we regress $\Delta diffVA_{it-1}$ and $\Delta diffVA_{it+1}$ on $\Delta countyVA_{it}$, while non-random moving predicts a non-zero coefficient.

For our "after" placebo test (i.e. observing changes in residuals from t to $t+1$), the students used in the student-switching experiment in Section 3 are all in middle school in $t+1$. American middle schools typically assign students to many teachers. Therefore, for the forward placebo test, we adjust a student's residuals relative to countymates in the same grade, rather than classmates.

A second challenge for the placebo test is that the "before" $(t-1)$ placebo test requires that the student be observed in $t-3$, so that $studentVA$ can be constructed in $t-2$, so that the change in student residuals (relative to peers) can be observed between $t-2$ and $t-1$. Because we do not observe students until 3rd grade, this is not possible for students who move in 5th grade, which we use as our primary sample in Section 3 in order to ensure matching to teachers. Instead, for our $t-1$ placebo test, we use students who are first observed in their new county in 6th grade.

Finally, to avoid the inclusion of the moving student in the measurement of county value-added, we report results measuring $\Delta countyVA$ with the years reversed. So, if a student moves from county j in $t-1$ to county j' in t, our right-hand side variable is $countyVA_{j't-1} - countyVA_{jt}$. The results are unaffected by using $\Delta countyVA$ without years reversed.

Results of the placebo test are shown in Table 10 separately by subject. The results show no evidence that the location decision is correlated with either a pre-trend or a post-trend in either subject.

Table 10: Change in student residuals

	Before $(t-1)$		After $(t+1)$	
	$\Delta DiffVA_M$	$\Delta DiffVA_R$	$\Delta DiffVA_M$	$\Delta DiffVA_R$
$\Delta CountyVA_M$	0.024		-0.006	
	(0.062)		(0.058)	
$\Delta CountyVA_R$		-0.072		-0.079
		(0.096)		(0.090)
N	15,151	14,978	13,866	13,735

** indicates $p < 0.01$. Robust standard errors in parentheses. M denotes math, R denotes reading. Outcome variables are $\Delta diffVA_{t-1}$ in the left column and $\Delta diffVA_{t+1}$ in the right column, where $\Delta diffVA_{t-1}$ is constructed from $\Delta studentVA_{t-1}$ using an adjustment for same-grade classmates' average residuals and $\Delta diffVA_{t+1}$ uses an adjustment for same-grade countymates' average residuals. The regressor is constructed using reverse-year averages. See text for details.

C Correlation with error term

In this appendix, we present a simple calculation showing what correlation between a county's effects on test scores and its unobserved determinants of income effects would be required to produce zero correlation between counties' effects on test scores and incomes. This calculation is referenced in Section 5.

For simplicity of exposition, we drop notation dividing students by parental income type. As before, let W_c be a county's effects on incomes. Additionally, let $T_c = \sum_{g=0}^{12} TS_{cg}$ be county c's total effects on test scores, such that $W_c^{TS} = R * T_c$ in the model for some constant R reflecting the return to test score production. Finally, as in Section 5, let U be the non-test score determinants of county effects on incomes. The structural equation for W is

$$W_c = R * T_c + U_c.$$

We will now calculate the approximate degree of correlation between T and U required to produce the patterns seen in the data.

Let k be the number such that $var(k * R * T) = var(W)$. For example, if R were the return to teacher quality, as we assume in the main results of the model in Section 5, then, according to the results in that section, k should be approximately equal to 1, since $var(W_c)$ is approximately equal to $var(W_c^{TS}) = var(R * T_c)$ at that value of R. If instead R were half of the return to teacher quality, then k would be 2.

From the structural equation above, we have

$$var(W) = var(RT) + var(U) + 2cov(RT, U).$$

By the definition of k, $var(W) = k^2 var(RT)$. Therefore we have that

$$k^2 var(RT) = var(RT) + var(U) + 2cov(RT, U). \quad (1)$$

Rearranging gives

$$cov(RT, U) = \frac{1}{2}\left[(k^2 - 1)var(RT) - var(U)\right]. \tag{2}$$

Since $corr(T, U) = corr(RT, U)$, this then yields

$$corr(T, U) = \frac{k^2 - 1}{2}\frac{\sigma_{RT}}{\sigma_U} - \frac{1}{2}\frac{\sigma_U}{\sigma_{RT}}, \tag{3}$$

where σ denotes the standard deviation of the subscripted variable.

Based on the results in Section 4, let us assume that $cov(T, W) = 0$. By the structural equation, this gives that

$$\begin{aligned} 0 &= cov(T, RT + U) \\ &= cov(T, RT) + cov(T, U) \\ &= Rvar(T) + cov(T, U). \end{aligned}$$

Therefore, multiplying by R,

$$\begin{aligned} 0 &= R^2 var(T) + Rcov(T, U) \\ &= var(RT) + cov(RT, U). \end{aligned}$$

Plugging this into the right-hand side of (1), we have that

$$k^2 var(RT) = var(U) + cov(RT, U).$$

Then, from (2), we have

$$k^2 var(RT) = var(U) + \frac{1}{2}\left[(k^2 - 1)var(RT) - var(U)\right],$$

which, multiplying through by 2 and rearranging terms, simplifies to

$$(k^2 + 1)var(RT) = var(U).$$

Taking the square root gives

$$\sqrt{k^2 + 1}\,\sigma_{RT} = \sigma_U.$$

Finally, plugging this into (3) gives

$$\begin{aligned} corr(T, U) &= \frac{k^2 - 1}{2\sqrt{k^2 + 1}} - \frac{1}{2}\sqrt{k^2 + 1} \\ &= \frac{1}{2\sqrt{k^2 + 1}}\left[k^2 - 1 - (k^2 + 1)\right] \\ &= \frac{-1}{\sqrt{k^2 + 1}}. \end{aligned} \tag{4}$$

For example, suppose that R were such that the variance of RT were equal to the variance of W, as our model in Section 5 suggests would be the case if the return to test scores produced by a county were equal the return to test scores produced by a teacher. Then $k = 1$. Plugging into (4) above, this means that the correlation between T and U would have to be $\approx -.707$.

Suppose the true value of R were only half as large as would be expected based on the return to teacher quality. Then $k = 2$ and the required correlation from above would be $\approx -.447$. At $k = 3$, the correlation is $\approx -.316$; at $k = 4$, the correlation is $\approx -.243$.

D Measurement error

Measurement error in the assignment of teachers to students could generate results in which teachers appear to change value-added as they move across counties. As an extreme example, if the same teacher ID were assigned to a randomly selected teacher in the old county and another randomly selected teacher in the new county, then the measured jump in teacher value-added should be roughly equal to the difference in average value-added by county.

Teachers are tracked longitudinally in the North Carolina data using employment records which are matched between years on the basis of Social Security number. However, the process of matching a personnel file to student performances is somewhat more complicated. There are two main types of measurement error which we consider as explanations for the findings of Section 6.

The first type of measurement error stems from the fact that the teachers listed in the North Carolina data are the teachers who proctored the standardized test, not necessarily the primary instructor. To address this potential source of bias, we use the fact that a more precise match of students to teachers is available after 2006 using course attendance files. We first estimate the same test score value-added model for the years after 2006, with the exception that free/reduced price lunch status is not included in the regression due to being unavailable in these years. Then, we estimate the quasi-experiment of teachers switching counties, restricting the sample to only those teachers whose students in the data are listed as matching to the teacher in the following three senses: (i) the teacher is listed as the exam proctor; (ii) the teacher is listed in the attendance files as the instructor for the student's math class; and (iii) the teacher is listed in the attendance files as the student's instructor in the greatest number of courses. Using this refined sample, we find very similar results to the results reported in Section 6. Furthermore, the rate at which teachers are successfully linked to their students in this way is not statistically different from the rate of links in the overall sample of teachers (i.e. including those who do not switch counties in the year in question) and only a couple of percentage points lower than for those teachers who switch schools within the same county. These differences are not large enough to drive our results. In particular, mismeasurement is not enough to

lead to any correlation between change in teacher value-added and change in school value-added for teachers who move *within* a county, so very small differences in the match rate cannot explain why we obtain such different results for *between*-county switchers.

A second type of mismeasurement arises from the possibility of an erroneous match of teacher names in personnel files to the names listed on testing forms. This possibility is discussed extensively in NCERDC documentation and we refer the reader to their documentation for more details.[41]

Two pieces of evidence suggest that neither form of measurement error drives our results. One is the placebo test shown in Section 6, showing that teachers' value-added moves somewhat in the *opposite* direction of the change in county value-added in the placebo years. If teacher identity were mismeasured, we would not expect to see this.[42]

Another test of mismeasurement is to look for heteroskedasticity in the regression of change in teacher value-added on change in county value-added. The reasoning is as follows. Suppose that some teachers are correctly labeled in the data and retain the same value-added as they move across counties (up to sampling error), while other observations of teacher switches in fact involve incorrect assignment of teachers to students. Then, when a teacher is correctly assigned to students in the data, the year-on-year change in a teacher's value-added would be merely noise. But if there is incorrect assignment of teacher to students in the new county, then the jump will be large whenever there is a substantial difference in value-added between the counties, and will be proportional to the difference in value-added between counties. This means that, if there is a mixture of correct and incorrect matches of teachers to students, we would expect to see a larger variance of the error term in the teacher-switching regression when the difference in county value-added is large. But tests for this form of heteroskedasticity fail to reject that the regression is homoskedastic.

E Returns to literacy and numeracy

As described in Section 7, one possible explanation for our findings is that counties' effects on test scores represent narrowly targeted increases in literacy and numeracy without broader gains in cognition, and that such narrow gains are not useful for incomes in adulthood. The primary reason we theorize these skills might not command a large return is that these skills are widespread due to universal education. When a skill is not scarce, it is unlikely to be rewarded with higher pay in the labor market. To support the claim that this hypothesis is at least plausible, in this appendix, we

[41] Documentation is available at https://childandfamilypolicy.duke.edu/research/nc-education-data-center/list-files-variables/.

[42] Even if high and low value-added counties systematically differed in their influence on teachers' trends in value-added, this would manifest itself as opposite-signed estimates for m positive versus for m negative.

document some empirical evidence suggesting that literacy and numeracy per se may not be strongly rewarded on the labor market.

We combine data from the Occupational Information Network (O*NET) with data from the 2010 American Community Survey (ACS). O*NET is a United States Department of Labor database which assigns numeric values to the skill requirements of occupations. Using O*NET data, we can assign skill levels to each individual in the ACS data based on their primary occupation. The ACS data also contains information on income. For all individuals born between 1980 and 1986 with non-missing data, we estimate the following regression:

$$ln(Inc)_i = \alpha_0 + X_i'\alpha_1 + \xi_i,$$

where X_i is a subset of the skills assigned to i's occupation by O*NET, standardized to have mean 0 and standard deviation of 1 in the sample. We first allow X to be the required level of Written Expression, which is defined by O*NET to mean "the ability to communicate information and ideas in writing so others will understand." We take this to be a narrow analogue of literacy skills. We next allow X to be the required level of Mathematical Reasoning, which is defined to mean "the ability to choose the right mathematical methods or formulas to solve a problem," which we interpret as a narrow analogue of numeracy skills. Then we construct a new variable, $Total$, by taking the sum of the standardized Mathematical Reasoning and Written Expression ratings, then standardizing the sum.

Columns 1 through 3 of Table 11 shows the results of these regressions. A one standard deviation increase in literacy, numeracy, and combined skills are each associated with 30% higher earnings.

However, this high apparent return might reflect other job requirements which are correlated with numeracy and literacy. In particular, simply controlling for the required level of Deductive Reasoning (defined as "the ability to apply general rules to specific problems to produce answers that make sense") virtually eliminates the estimated return to Mathematical Reasoning, and entirely eliminates the estimated return to Written Expression. This suggests that the apparent return to literacy and numeracy may not be driven by facility with numbers and writing themselves, but rather by the correlation between their use and job requirements for abstract reasoning abilities.

It may still be possible that literacy and numeracy per se are substantially rewarded in the labor market if, conditional on deductive reasoning requirements, occupational requirements for these skills are negatively correlated with other skill requirements.[43] We emphasize that the regressions above are far from conclusive.

Nonetheless, this evidence at least suggests the *possibility* that literacy and numeracy per se do not command large wage premia. Instead, the broader capability to engage in reasoning might be more essential. If that is the case, interventions

[43]For example, regressions of log of income on physical skills tend to produce negative coefficients, presumably because jobs which require physical skill typically do not require cognitive skills.

Table 11: Log income on skill requirements

	(1) $ln(Inc)$	(2) $ln(Inc)$	(3) $ln(Inc)$	(4) $ln(Inc)$	(5) $ln(Inc)$	(6) $ln(Inc)$
Written Expression	.300**			-.007		
	(.002)			(.004)		
Math Reasoning		.300**			.036**	
		(.002)			(.004)	
$Total$			.314**			.025**
			(.002)			(.005)
Deductive Reasoning				.347**	.311**	.318**
				(.004)	(.004)	(.005)
R^2	.170	.171	.187	.221	.221	.220
N	113,164	113,164	113,164	113,164	113,164	113,164

** indicates $p < .01$. Robust standard errors in parentheses.

which narrowly promote literacy and numeracy without cultivating broader cognitive function may not be beneficial to students' labor market prospects.