Preliminary: Please do not cite without permission

Teacher Effects on Achievement and Anthropometrics:

A Cautionary Tale

June 2011

Marianne Bitler

Department of Economics, UC Irvine & NBER

Thurston Domina

Department of Education, UC Irvine

Emily Penner

Department of Education, UC Irvine

We thank Greg Duncan, Jim Wyckoff, Dean Jolliffe, Richard Buddin, George Farkas, Jesse Rothstein, Sean Reardon, Michal Kurleander, and Marianne Page for helpful comments.

In the last several years, teachers have come under increasing scrutiny in the search for strategies to improve the quality of American public education and narrow inequalities in educational outcomes. The growing availability of data that links students to their teachers and measures student achievement from year to year has made it possible to estimate the contributions that teachers make to student achievement. As a result, researchers have investigated teacher value-added using administrative data from North Carolina (Clotfelter et al. 2006, Goldhaber 2007, Rothstein 2011), Texas (Rivkin et al. 2005), New York (Kane, Rockoff, & Staiger 2006), and Florida (Harris & Sass 2006), as well as data from several large urban school districts, including Chicago (Aaronson, Barrow & Sander 2007), Los Angeles (Buddin & Zamarro 2008, Buddin 2011, Kane & Staiger 2008), and San Diego (Koedel & Betts 2007, 2011). Investigations into the extent that teachers vary in their effectiveness have yielded estimates ranging from 0.08 to 0.85 standard deviations. However, recent value-added studies consistently demonstrate wide variance in teacher effectiveness, such that a one standard deviation difference in teacher effectiveness is associated with 0.15-0.30 standard deviation difference in student achievement. This variation in teacher effectiveness suggests that a student who has an effective teacher will experience nearly a full year's more achievement growth than a student who has an ineffective teacher.

Based on these findings, a consensus has emerged among school reformers and educational policy-makers that teacher quality is the most important school-based input in student achievement. In a March 2010 speech on education policy, President Barack Obama encapsulated this consensus, arguing: "From the moment students enter a school, the most important factor in their success is not the color of their skin or the income of their parents, it's

the person standing at the front of the classroom." To improve teacher quality, the White House is pressuring states to measure teacher value-added and implement teacher hiring and compensation systems that reward effective teachers.

Despite growing acceptance of the importance of studying teacher effectiveness and the embrace of value-added by policy makers (McCaffrey et al. 2004), there is some evidence to suggest that value-added models provide biased estimates of teacher effects and overstate the extent to which teachers differ from one another (Ladd 2008; Rothstein 2009, 2010; Kane and Staiger 2002; Aaronson, Barrow, and Sander 2007). Most notably, Rothstein constructs a falsification test for the most commonly used value-added models, investigating the measured impact of fifth grade teachers on fourth grade math and reading score gains.[1] While the notion that teachers should be able to causally influence student achievement *before* students enter their classrooms is implausible, Rothstein's value-added models show that these implausible teacher effects exist, suggesting that bias from unobservables impacts these estimates of teacher's influences on student test scores, leading value-added models to exaggerate the variation in teacher effectiveness.

The current study proposes an additional falsification test of value-added models of teacher effectiveness. Using a nationally-representative data set, the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K), we estimate the effects of Kindergarten teachers on student math and reading achievement as well as on two biological variables, height and weight, at the end of the Kindergarten year. We then compare the distribution of teacher effects on students' math and reading achievement with the distribution of teacher effects on height and weight. Our premise is that we should not find statistically significant teacher effects

[1] Kane and Staiger (2008) test for bias associated with some typical value-added models in an experimental setting, finding little evidence of bias. Koedel and Betts (2011) test for bias in value-added models with data from the San Diego school district, and find that with sufficiently rich controls, bias is not a problem.

on height or weight unless height and weight are correlated with achievement, regardless of how the teacher effects on achievement arise. At a minimum, assuming that teachers do not affect child nutrition enough to influence child height, any teacher effect in height in a developed country like the US cannot be causal but must reflect the correlation of height with other child characteristics.[2] Therefore, evidence of significant variation across teachers in gains in height, or to a lesser extent in weight, should raise some question about the usefulness of value-added estimates of teacher quality.

In addition, we compare measured teacher effects on math and reading with simulated teacher effects generated by randomly assigning ECLS-K students to Kindergarten teachers, testing a null of no teacher impact. Our analyses reveal substantial variation in teacher effects on height and weight. Further, we find that even when students are randomly assigned to teachers, teacher value-added models suggest that teachers have differential effects on student achievement. While these placebo teacher effects are not as large as the more plausible teacher effects in math and reading, our analyses suggest that teacher value-added models may overstate the extent to which teachers differ from one another in contemporary American public schools.

Participants

This study uses data from the Early Childhood Longitudinal Survey, Kindergarten Cohort (ECLS-K), a nationally representative longitudinal study that follows students from the 1998

[2] Obviously, in less developed countries with shortages of sufficient food and micronutrients, height and weight are correlated with child development. An investigation of the impact of food insecurity in the US using the ECLS-K found it to have no effect on height and weight levels or growth in kindergarten (Winicki and Jemison 2003). There is also a literature linking adult wages to height even in developed countries (e.g., see the literature review in Persico, Postlewaite, & Silverman 2004). However, in the absence of widespread starvation, there is little evidence linking wages to height in childhood. For example, Persico, Postlewaite, & Silverman find that teen height but not adult or child height is correlated with adult wages. Further, our estimates of teacher value-added use changes in height (or weight or achievement) as the dependent variables, or alternatively, use the level of height (or other outcomes) as the dependent variable but control for lagged height (lagged outcomes). It is less plausible that changes in height are correlated with unobserved ability, and empirically we demonstrate below that changes in height and weight are uncorrelated with changes in achievement.

kindergarten class. The ECLS-K study was administered by the Department of Education and sampled over 20,000 children entering Kindergarten in the fall of 1998 in roughly 1,000 schools, with an average of twenty-three students per school. Students were sampled from both public and private schools from sampling units (a county or a group of counties) across the country. The initial data collected on the students were gathered during the fall of Kindergarten, with subsequent collection in the spring of Kindergarten, the fall and spring of first grade, and the spring of third, fifth, and eighth grades.

This study uses data from the fall and spring of Kindergarten as students become too dispersed in classrooms in later grades to be used for this type of analysis. The benefit of examining Kindergarten exclusively is that there is less room for student sorting into classrooms to be influenced by teacher and principal intervention, which is of concern in later grades (Rothstein 2009). Thus to the extent that non-random sorting of students is a problem in typical value-added models, it should be less of an issue in Kindergarten. Because of the extensive teacher and parent interviews conducted by the ECLS-K, the data allow for controls of family background characteristics that may influence student sorting into classrooms on the basis of parental influence. The preliminary analyses reported here do not exploit these controls; but future analyses will.

The sample is restricted to include only first-time Kindergarten students in public schools with the same teacher in fall and spring of Kindergarten, in classrooms in which five or more students were included in the ECLS-K sample. A comparison of Kindergarten students from the full sample of public schools and the study sample is presented in Table 1. From Table 1, it is clear that our sample inclusion criterion requiring 5 children per teacher combined with the sample design (around 20 children sampled per school, randomly across teachers) has caused our

sample to have more children per teacher, and fewer teachers per school than the full sample of public school children. On individual characteristics, our sample is slightly less likely to be Hispanic and to speak another language at home than the full sample. Our sample has children whose parents have slightly more completed education than the full sample. On net, the samples are very similar.

[Insert Table 1 Here]

The structure of the ECLS-K presents several challenges for the estimation of teacher value-added models, and with one exception (Jennings and DiPrete 2010), these data have not been widely used in the teacher effects literature. Since the study is based on a sample of students within schools, we only have data for a subset of the students in a given teacher's classroom. Furthermore, the ECLS-K data are limited in that they follow a single cohort of students and are thus unable to examine a teacher's effect on student growth with multiple groups of students. Estimates of teacher effects generated based on the experience of large numbers of students over several years are more precise than estimates of teacher effects based on the experience of a small number of students over one year (Buddin 2011). As a result, our teacher value-added estimates are likely less precise than estimates generated using administrative data. However, the ECLS-K is otherwise well-suited to our analyses in that it provides measures of student biological growth during the school year. We know of no other survey or administrative data that currently provide measures of student biological growth along with measures of student achievement for multiple children in classrooms. An additional advantage to using the ECLS-K data is that it spans the whole country, and examines children with tests that were intended to have no ceiling effects. Furthermore, since any biases that result from the complex sample design are equally likely to appear in our estimates of teacher effects

on math and reading as in our estimates of placebo teacher effects, we suspect that our findings can be replicated using administrative data that link teachers and students and provide information on student achievement gains as well as biological growth.

Measures

ECLS-K test score and anthropometric data were collected by trained assessors, supervised by trained field supervisors. Assessors received roughly 30 hours of in-person training in August and September of 1998, preceded by eight hours of home study training.

Dependent Variables: Academic and Anthropometric Growth

The present analysis uses academic and anthropometric growth from fall to spring of Kindergarten as its outcomes of interest. Academic growth is measured using IRT scores for reading and math. The IRT instruments were designed particularly for the ECLS-K, but were based on existing instruments, including Children's Cognitive Battery (CCB), Peabody Individual Achievement Test—Revised (PIAT-R); Peabody Picture Vocabulary Test—3 (PPVT-3); Primary Test of Cognitive Skills (PTCS); and Woodcock-Johnson Psycho-Educational Battery—Revised (WJ-R). The mathematics and reading IRT scores were collected in one-on-one assessment sessions with the child by the trained assessors. Both the reading and the mathematics assessments evaluate students across a number of content strands using adaptive testing methods that allowed the tests to be tailored to the students and helped prevent test score ceilings by allowing students to continue to more advanced questions. Students were not asked to write anything or explain their reasoning, but were instead asked to point or use verbal responses. The data were collected using computer-assisted interviewing technology. The reading assessment covered several content strands, including initial understanding, developing interpretation, personal reflection and response, and demonstrating a critical stance. The

mathematics assessment evaluated students' knowledge of number sense, properties and operations; measurement; geometry and spatial sense; data analysis; statistics; and probability, patterns, algebra, and functions. The reported reliabilities of the reading IRT scores for fall and spring of K are 0.93 and 0.95 respectively. The reported reliabilities of the math IRT scores for fall and spring of K are .92 and .94.

Anthropometric growth is measured using height, in inches, and weight, in pounds. In the fall and spring of Kindergarten, height and weight were measured as part of the one-on-one child assessments. Height was collected using a Shorr Board and weight was collected using a bathroom scale. Height and weight were both taken twice to prevent error and provide an accurate reading. For height, if the two measurements were less than two inches apart, the average of the two measurements was used as the composite measurement of height. If the measurements were more than two inches apart, the measurement closest to 43 inches (average height of a five year-old child) was used as the composite. Similarly, if the two measurements of the child's weight were less than five pounds apart, the average of the two measurements was used as the composite measurement of weight. Otherwise, the measurement closest to 40 pounds (the average weight of a five-year old) was used as the composite.[3]

From the measurements of reading, math, height and weight provided in the ECLS-K dataset, scores were standardized and the fall score was subtracted from the spring score. From the differenced standardized scores, the top and bottom five percent for each of the academic and anthropometric change scores were truncated. Additional specifications instead use spring measures as the dependent variable, controlling for the fall measures; also after truncating the top and bottom 5 percent of both measures. Means, standard deviations, and the amount of change

[3] The bulk of height measurements are even quarters or tenths of an inch, suggesting either measurements that were an even multiple of half inches or two tenths of inches apart, or widespread issues with one measurement.

from fall to spring of the five dependent variables are presented in Table 2. Recall that all of these are either standardized versions of the outcomes (z-scores), or differences in these standardized z-scores.

[Insert Table 2 Here]

As laid out above, our falsification test of teacher value-added models is based on the assumption that teachers cannot influence students' height and weight growth, and thus, that estimated teacher effects on these anthropometric measures should not differ significantly from zero. We explore this by looking at the correlation between the differences in scores and spring scores, presenting the correlations in Table 3. Consistent with this assumption, Table 3 indicates that there is hardly any correlation between student achievement and student height and weight. Student scores on reading and math correlate at 0.60 and student height and weight correlate at 0.59. However, correlations between achievement measures and anthropometric measures are all considerably less than 0.10. The second panel of Table 3 further reveals that the correlations between student test score growth and student biological growth are nearly zero.

[Insert Table 3 Here]

Methods

The analyses reported here are based upon two simplified versions of the value-added models that are typical in the teacher effects literature. First, we estimate a simple gain score model of teacher value-added:

$$Y_{ij,\text{Spring K}} - Y_{ij,\text{Fall K}} = \sum \beta_j + \varepsilon_{ij}$$

where the dependent variable is the difference between test scores or anthropometric measurements for student *i* in classroom *j* in the Spring of Kindergarten and their corresponding

value from the previous Fall and the parameter of interest, $\sum \beta_j$ is a matrix of teacher fixed effects. [4]

Second, we estimate a lagged dependent variable model of teacher value-added, which only differs from the gain score model by moving Fall Kindergarten measurements to the right hand side of the equation:

$$Y_{ij,\text{Spring K}} = Y_{ij,\ \text{Fall K}} + \sum \beta_j + \varepsilon_{ij}$$

In both models, the key statistic of interest is the standard deviation of the teacher fixed effects. Other models in the teacher value-added literature include contextual controls for students, teachers, and schools. If the data permit, many models also include student and school fixed effects. Preliminary analyses indicate that our findings are not model-sensitive; nonetheless, we plan include extensive student controls as well as school fixed effects in future drafts. Following Aaronson, Barrow, and Sander (2007), we present both the unadjusted standard deviation of the teacher fixed effects as well as an adjusted standard deviation which subtracts off an estimate of the sampling variation in the teacher fixed effect-measures as the average squared standard error of the teacher fixed effect. The ECLS-K is a complex sample survey; our main estimates use the child kindergarten panel year weight. Our results currently present models with robust standard errors.[5] For each outcome, we use the full sample of observations with data on the outcome of interest. In addition to the unadjusted and adjusted standard deviations, we report the results of F-tests that the teacher fixed effects are jointly zero and also display figures of the estimated teacher fixed effects, normalized to be mean zero. We also present figures of the

[4] Future drafts will consider random effects models as well.
[5] So far we have explored adjusting the standard errors of the teacher fixed effects by jackknifing. Later work will adjust for the complex sample nature more completely, by imposing a null of no systematic teacher effects by randomly assigning the dependent variable to children from the sample, and bootstrapping these null estimates (following Abadie, 2002), while following the stratification and clustering of the original sample.

estimated teacher fixed effects for the randomly assigned dependent variables. Future work will test whether these distributions are different, and will also explore the extent to which the estimates are sensitive to controlling for the kind of demographic controls often in administrative data, for school fixed effects, and for the kitchen sink of family characteristics that are particular to the ECLS-K. We will also explore the impact of altering the number of students required for each teacher to appear in the sample.

Results

Table 4 summarizes the observed variation in teacher fixed effects on Kindergartener's gains in math and reading skills, as well as growth in height and weight over the Kindergarten year. Our estimates of the variation in teacher effects on student learning are roughly in line with the estimates of teacher effects in that appear elsewhere in the literature. The F-tests reported in the first two columns of Table 4 indicate that teacher fixed effects in both math and reading are both jointly significantly different from zero. The standard deviation of the estimated teacher fixed effects is 0.360, while a change from the 25^{th} to 75^{th} percentile of teacher effectiveness results in a change of 0.430. The corresponding changes for reading are a standard deviation of 0.370, with a change from the 25^{th} to 75^{th} percentiles of the teacher effectiveness distribution resulting in a change of 0.508.[6][7] However, since there are fewer students per teacher in the ECLS-K than in the administrative data files used elsewhere in the literature, we find that

[6] Since we have adjusted the dependent variables to be z-scores, the standard deviation of the untrimmed scores should be 1. The standard deviation of the gain measures before trimming ranges from 0.95-0.98.

[7] Our estimates of the variation of teacher effects in math and reading in the ECLS-K data are somewhat larger than Jennings and DiPrete's (2010) estimates using the same data. The primary difference is methodological. Their principal estimation strategy relies on random effects from a three-level hierarchical model; while our estimation strategy relies on fixed effects estimates. Future drafts will consider random effects models as well. An additional potential source of discrepancy lies in the sample restrictions. They restricted their sample to first time kindergarteners, in the same school in fall and spring of kindergarten, at schools with two or more kindergarten teachers, in classes shared with two or more other students. While their analyses were not sensitive to the number of students per classroom (they tried additional models requiring five students per classroom), it is possible that the discrepant restrictions also contribute to different teacher effect estimates.

adjusting the standard deviation of teacher fixed effects for sampling error shrinks the estimated variation of teacher effects by roughly 30-40%. The figures reported in the second row of Table 4 indicate, therefore, that the adjusted standard deviation of teacher fixed effects is 0.23 in math and 0.27 in reading. These estimates of the variation in teacher effectiveness are somewhat smaller than the estimates available elsewhere in the literature. However, this is perhaps not surprising, since the Kindergarten teachers that we study often spend just half a school day with their students. Additionally, since there are fewer observations contributing to each teacher fixed effect, one might also expect the correction to the standard deviation for sampling error to be larger than if we had data on all the students, leading to a bigger difference between the unadjusted and adjusted standard deviations.[8]

[Insert Table 4 Here]

The findings reported in the third and fourth columns of Table 4 are more surprising. While it is not plausible that teachers in the US can influence student biological growth, we find evidence that teachers differ significantly from one another in their observed effect on student height and weight. The unadjusted standard deviation of teacher fixed effects is 0.21 in height and 0.25 in weight. As in math and reading, we find that the adjusted standard deviations are smaller, but still different from zero at 0.15 in height and 0.14 in weight. In either the unadjusted or the adjusted case, the estimated variation in teacher effects on height is approximately 60 percent the size of the estimated variation in teacher effects on math and reading.

Figures 1 through 4 provide a more detailed look at the relationships among estimated teacher effects in math, reading, height and weight by graphing the distribution of teacher fixed

[8] Suppose the teacher fixed effect estimates were left relatively unchanged by adding more observations per teacher. Then the unadjusted standard deviation would be very close to the one here. Presumably, the average squared standard error of the fixed effects would be smaller because the teacher fixed effects would be estimated more precisely. Thus, the adjusted standard deviations would likely be larger.

effects resulting from these models. Consistent with the findings reported in Table 3, we find that teacher effects on math and reading are fairly evenly distributed around zero. While the measured teacher effects on height and weight are much more tightly concentrated around zero, there are a non-trivial number of teachers on both positive and negative tails. A Kolmogorov-Smirnov test that the distributions are normal fails to reject the null of normality for reading or weight, but rejects the null that the distribution is normal for math and height. We also tested for normality with a ShapiroWilk test, and here the null of normality is rejected for reading, math, height and weight.

[Insert Figures 1-4 Here]

Table 5 reports the results of of teacher value-added, controlling for lagged test scores. Consistent with the results reported in Table 4, we find evidence of substantial teacher effects on height and weight. The estimated variation in these teacher effects are 30-60 percent the size of the estimated variation in teacher effects on reading and math.

In Figures 5 through 8, we graph the distribution of teacher fixed effects estimated from lagged models. Again, these graphs indicate that measured teacher effects on math and reading are more widely dispersed than measured teacher effects on the anthropometric placebos. Despite the implausibility of teachers causally influencing their students' height and weight, however, the lagged value-added models in Figures 7, and 8 give the impression that a substantial number of teachers do exert such influence on height and weight.

[Insert Table 5 and Figures 5-8 Here]

Finally, Figures 9-12 report early results from a second falsification test, in which we estimate teacher value-added models on randomized student gain scores. Like the models reported in Table 4 and graphed in Figures 1-4, these models estimate teacher fixed effects on

student achievement and anthropometric gains. Unlike those earlier models, however, student values on the dependent variables in these models are randomly assigned in order to provide another indication of what true random noise looks like in teacher fixed effects estimates. Each child is assigned a randomly selected gain score, and then the teacher fixed effects are estimated. As expected, these randomized teacher effects estimates are uncorrelated with the observed estimates of teacher effects reported above.[9] We test whether the distribution of fixed effects from the randomly assigned gains is the same as the distribution of teacher fixed effects from the real data, using a Kolmogorov-Smirnov test.

The distribution of teacher fixed effects for the random gains in Figures 9-12 still looks as if it has fatter tails in math and reading than in height and weight. This may be because the standard deviation in our trimmed values for math and reading gains is wider than the standard deviation in height and weight gains for the students in the sample. However these randomized teacher fixed effects are clustered more closely around zero than each of the non-randomized teacher fixed effects reported in Figures 1 through 4. Kolmogorov-Smirnov tests overwhelmingly reject that the distributions for teacher effects for the randomly assigned gains are the same as the distributions of teacher effects for the true data for all 4 outcomes. Similarly, Shapiro-Wilk tests strongly indicate that the distribution of teacher effects on randomly assigned gains is not normal.

Conclusions

Efforts at reforming K-12 education in the use over the last decade have led to improved data systems and a body of research showing strong links between teacher effectiveness and student achievement.

[9] The correlation between the randomly assigned gain teacher effects and real teacher effects is 0.013 on math; -0.004 on reading; 0.005 on height; and 0.008 on weight.

This has led to a consensus among school reformers and educational policy makers that teacher quality is a key input for learning. In this paper, we estimate several of the most commonly used value-added models using data from the 1998 Early Childhood Longitudinal Study-Kindergarten. We first confirm that our estimates of teacher value-added models of math and reading predict similar variation among teachers to the existing literature. We then estimate teacher value-added models on height and weight, motivated by the expectation that these anthropometric outcomes, height in particular, can hardly be affected by teacher effectiveness in a developed economy. We find preliminary evidence suggesting that the adjusted and unadjusted standard deviations in teacher fixed effects for height and weight are at least half the magnitude of those for math and reading, while their distributions are more tightly concentrated around their means. These findings suggest that standard deviations of teacher value-added estimates may overstate the extent to which teachers differ from one another in their educational effectiveness.

References

Aaronson, D., L. Barrow, and W. Sander. 2007. "Teachers and student achievement in the Chicago public high schools." *Journal of Labor Economics* 25: 95-135.

Abadie, A. 2002. "Bootstrap tests for distributional treatments effects in instrumental variable models." *Journal of the American Statistical Association* 97: 284-292.

Buddin, R., and G. Zamarro. 2009. "Teacher qualifications and student achievement in urban elementary schools." *Journal of Urban Economics* 66:103-115.

Clotfelter, C. T., E. J. Glennie, H. F. Ladd, and J. L. Vigdor. 2008. "Teacher bonuses and teacher retention in low-performing schools: Evidence from the North Carolina $1,800 Teacher Bonus Program." *Public Finance Review* 36:63.

Goldhaber, D. 2007. "Everyone's doing it, but what does teacher testing tell us about teacher effectiveness?" *Journal of Human Resources* 42: 765-794.

Hanushek, E. A. 1992. "The trade-off between child quantity and quality." *Journal of Political Economy* 84-117.

Harris, D. N., and T. R. Sass. 2006. "Value-added models and the measurement of teacher quality." *Unpublished manuscript*.

Jennings, J. L., and T. A. DiPrete. 2010. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology of Education* 83:135.

Kane, T., J. Rockoff, and D. O. Staiger. 2006. *What Does Teacher Certification Tell Us About Teacher Effectiveness? Evidence from New York City"*. New York: NBER Working Paper No. 12155.

Kane, T. J., and D. O. Staiger. 2008. "Estimating teacher impacts on student achievement: An experimental evaluation."

Kane, T. J., and D. O. Staiger. 2002. "The promise and pitfalls of using imprecise school accountability measures." *Journal of Economic Perspectives* 91-114.

Koedel, C., and J. Betts. 2007. "Teacher quality and educational production in secondary school."

Koedel, C., and J. R Betts. 2011. "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique." *Education Finance and Policy* 1-26.

Ladd, H. F. 2008. "Teacher effects: What do we know?" Northwestern University, Chicago, Illinois.

McCaffrey, D. F., J. R. Lockwood, D. Koretz, T. A. Louis, and L. Hamilton. 2004. "Models for

value-added modeling of teacher effects." *Journal of Educational and Behavioral Statistics* 29:67.

Persico, N., A. Postlewaite, and D. Silverman. 2004. "The effect of adolescent experience on labor market outcomes: The case of height." *The Journal of Political Economy* 112:5, 1019-1053.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica* 73:417-458.

Rothstein, J. 2009. "Student sorting and bias in value-added estimation: Selection on observables and unobservables." *Education Finance and Policy* 4:537-571.

Rothstein, J. 2010. "Teacher quality in educational production: Tracking, decay, and student achievement." *Quarterly Journal of Economics* 125:175-214.

Winicki, J., and K. Jemison. 2003. "Food insecurity and hunger in the kindergarten classroom: its effect on learning and growth." *Contemporary Economic Policy* 21(2):145-157.

Table 1: Summary statistics, full sample in public school and our sample

	Full sample	Our sample
Male	0.52	0.51
Black	0.17	0.17
Hispanic	0.20	0.15
Asian/Pacific Islander	0.03	0.03
Hawaiian/other PI	0.01	0.004
American Indian	0.02	0.02
Multiple Races	0.02	0.02
Speak other language at home	0.22	0.19
One parent family	0.25	0.24
First time in Kindergarten	0.96	1
Sampled children per teacher	8.35	10.16
Sampled teachers per school	4.10	3.09
Parent's education:		
High school dropout	0.18	0.14
AA degree	0.06	0.06
College, no degree	0.25	0.27
BA	0.13	0.15
MA/MS or more education	0.04	0.05
N	14,359	9,168

Table presents weighted summary statistics for the full sample of children in public school and for observations satisfying our eventual sample restrictions. All observations required to have a non zero first year weight (BYCW0). Our sample for each specification is first time Kindergarten students in public school with the same teacher in fall and spring of Kindergarten, who share the teacher with at least 4 other students, all of whom have a value for the dependent variable. Since these samples differ slightly, we present summary statistics here for first time Kindergarten students with the same fall/spring teacher sharing that teacher with at least 4 other students, who have a non-missing value for at least one of the dependent variables. N is maximum number of observations for which the relevant variables are defined.

Table 2: Summary statistics, dependent variables

	Mean	Std. Dev.
A: Trimmed fall z-scores		
Math	-0.044	0.689
Reading	-0.086	0.601
Height	0.018	0.744
Weight	-0.052	0.651
B: Trimmed spring z-scores		
Math	-0.057	0.698
Reading	-0.108	0.563
Height	0.012	0.738
Weight	-0.054	0.661
C: Trimmed z-score of difference		
Math	-0.065	0.719
Reading	-0.104	0.658
Height	-0.007	0.403
Weight	-0.032	0.534

Table presents weighted summary statistics for our dependent variables. Variables in panel A are fall z-scores, with top and bottom 5% trimmed. Variables in panel B are spring z-scores, with top and bottom 5% trimmed. Variables in panel C are the z-score of the difference in the measures, with top and bottom 5% trimmed. All required to have a non zero first year weight (BYCW0). Sample for each specification is first time Kindergarten students in public school with the same teacher in fall and spring of Kindergarten, who share the teacher with at least 4 other students, all of whom have a value for the dependent variable.

Table 3: Correlation in dependent variables

A: Trimmed spring z-scores				
	Math	Reading	Height	Weight
Reading	0.597			
Height	0.085	0.046		
Weight	0.009	-0.013	0.592	
B: Trimmed z-score of difference				
	Math	Reading	Height	Weight
Reading	0.264			
Height	0.062	0.082		
Weight	0.025	0.048	0.196	

Table presents weighted correlations for the dependent variables. Panel A presents correlations for spring trimmed z-scores, panel B presents correlations for the trimmed z-score of the difference in measures. All required to have a non zero first year weight (BYCW0). Sample for each specification is first time Kindergarten students in public school with the same teacher in fall and spring of Kindergarten, who share the teacher with at least 4 other students, all of whom have a value for the dependent variable.

Table 4: Characteristics of the teacher fixed effects, dependent variable is z-score of gain, trimming top and bottom 5 percent, no other controls

	Math	Reading	Height	Weight
Unadjusted standard deviation	0.3604	0.3668	0.2144	0.2536
Adjusted standard deviation	0.2306	0.2728	0.1510	0.1437
F-test, teacher FE jointly 0	2.02	2.91	2.40	1.96
(p-value)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of teachers	795	771	838	822
Number of students	6211	6049	5672	6374

Table presents results of regressions predicting gains in various scores/anthropometric measures using ECLS-K data as a function of teacher fixed effects. All results are for the sample of children where at least 5 children had the relevant test score in each class, the children had the same first and second semester Kindergarten teacher, it was their first time in Kindergarten, and the panel first year child weight was positive. Each column presents results for one test score gain, where the dependent variable is the difference in the z-score, trimming the top and bottom 5% and the only controls are teacher fixed effects. The regressions use the child panel weight for the first year; the standard deviations are weighted by the number of children in the class. The first column presents the results for math scores, the second for reading the third for height, and the fourth for weight. The first row presents the unadjusted standard deviation of the teacher fixed effects, the second the adjusted standard deviation following Aaronson, Barrow, and Sander (2007), where the adjustment subtracts off the average squared standard error of the teacher fixed effect. The third row presents the F-statistics and fourth the p-value for the joint test that the teacher fixed effects are significantly different from zero. The number of teachers and students are in the fifth and sixth rows.

Table 5: Characteristics of the teacher fixed effects, dependent variable is spring z-score, trimming top and bottom 5 percent, controls for fall score

	Math	Reading	Height	Weight
Unadjusted standard deviation	0.2481	0.2401	0.1597	0.1153
Adjusted standard deviation	0.1659	0.1900	0.0959	0.0580
F-test, teacher FE jointly 0	2.24	3.17	1.78	1.68
(p-value)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of teachers	759	717	768	778
Number of students	5858	5576	5893	6031
Coefficient on fall value	0.727	0.631	0.899	0.941
(SE)	(0.010)	(0.010)	(0.006)	(0.005)

Table presents results of regressions predicting gains in various scores/anthropometric measures using ECLS-K data as a function of teacher fixed effects. All results are for the sample of children where at least 5 children had the relevant test score in each class, the children had the same first and second semester Kindergarten teacher, it was their first time in Kindergarten, and the panel first year child weight was positive. Each column presents results for one test score gain, where the dependent variable is the difference in the z-score, trimming the top and bottom 5% and the only controls are teacher fixed effects. The regressions use the child panel weight for the first year; the standard deviations are weighted by the number of children in the class. The first column presents the results for math scores, the second for reading the third for height, and the fourth for weight. The first row presents the unadjusted standard deviation of the teacher fixed effects, the second the adjusted standard deviation following Aaronson, Barrow, and Sander (2007), where the adjustment subtracts off the average squared standard error of the teacher fixed effect. The third row presents the F-statistics and fourth the p-value for the joint test that the teacher fixed effects are significantly different from zero. The number of teachers and students are in the fifth and sixth rows.

Figure 1: Teacher fixed effects, difference in z-scores in math, top and bottom 5% trimmed, no other controls

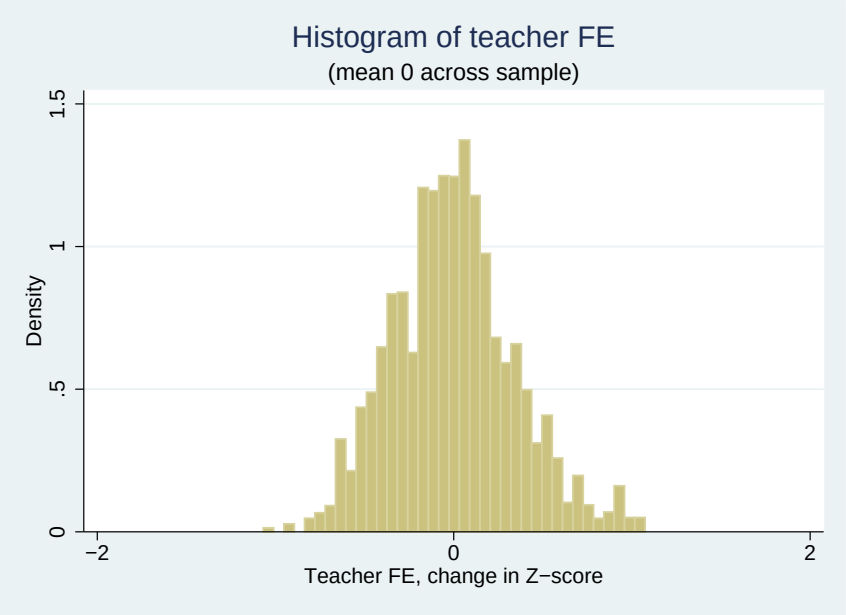

Figure 2: Teacher fixed effects, difference in z-scores in reading, top and bottom 5% trimmed, no other controls

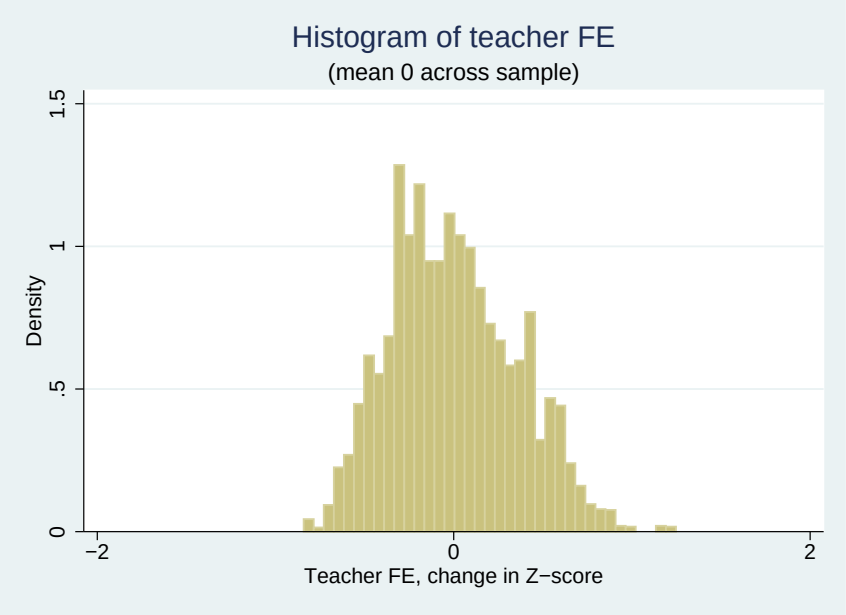

Figure 3: Teacher fixed effects, difference in z-scores in height, top and bottom 5% trimmed, no other controls

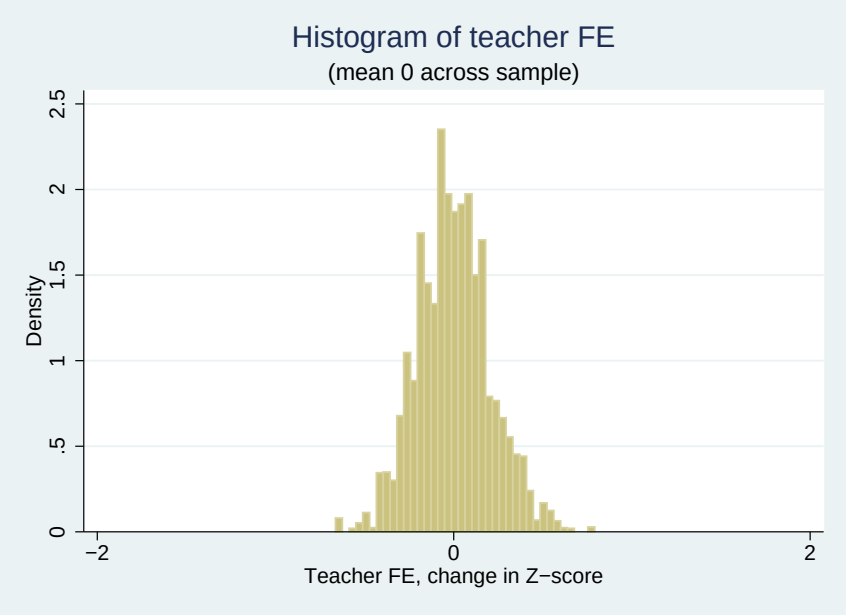

Figure 4: Teacher fixed effects, difference in z-scores in weight, top and bottom 5% trimmed, no other controls

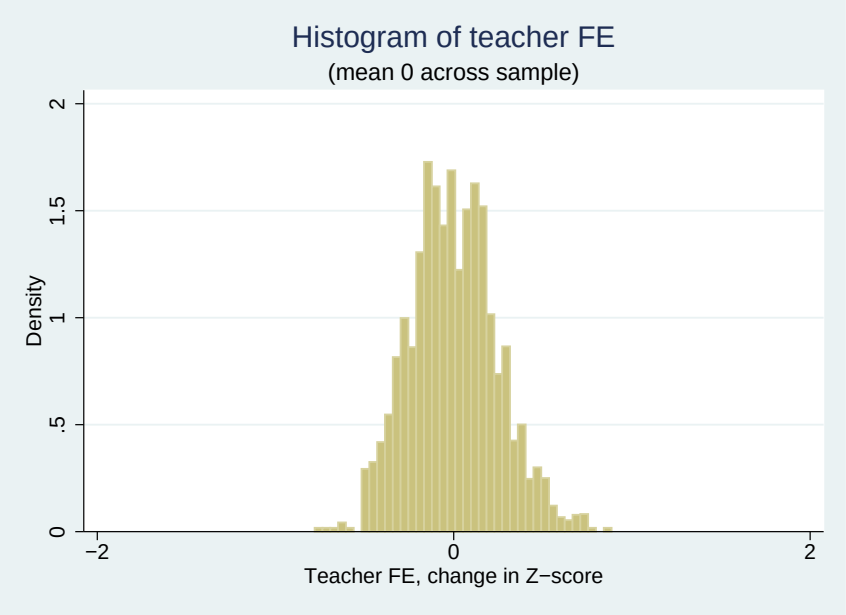

Figure 5: Teacher fixed effects, z-scores in spring math, top and bottom 5% trimmed, controls for fall score

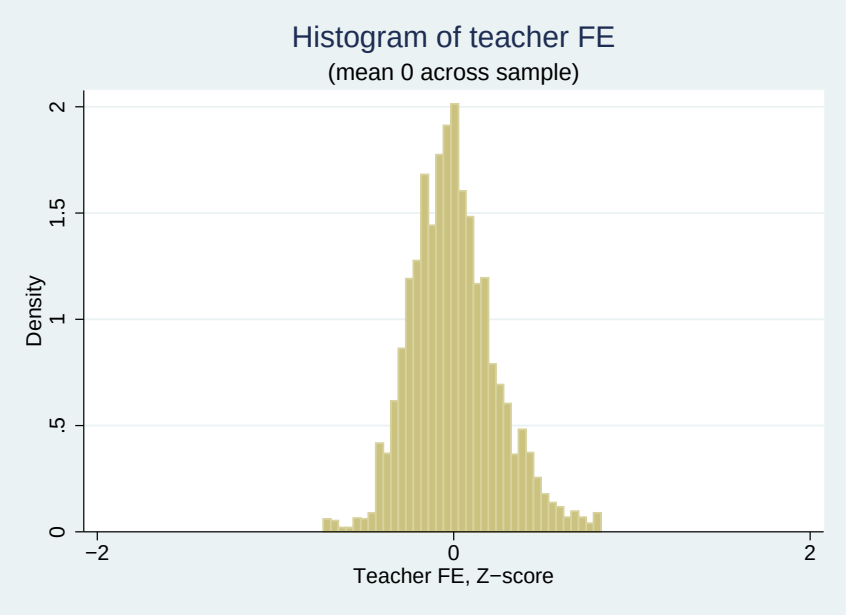

Figure 6: Teacher fixed effects, z-scores in spring reading, top and bottom 5% trimmed, controls for fall score

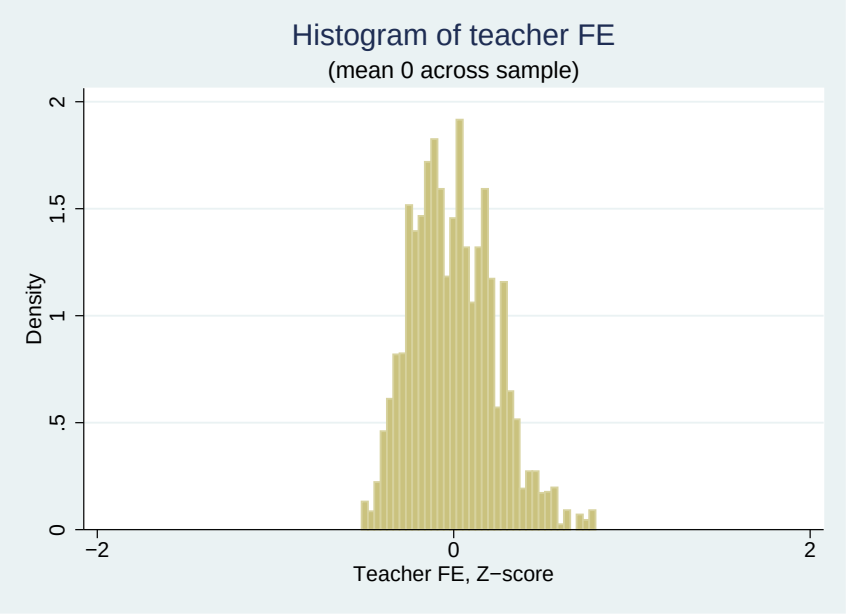

Figure 7: Teacher fixed effects, z-scores in spring height, top and bottom 5% trimmed, controls for fall score

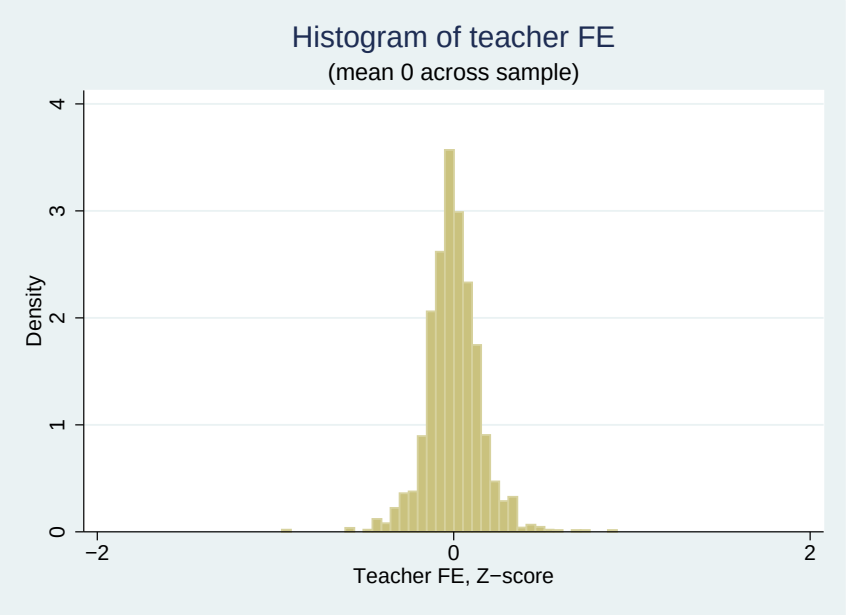

Figure 8: Teacher fixed effects, z-scores in spring weight, top and bottom 5% trimmed, controls for fall score

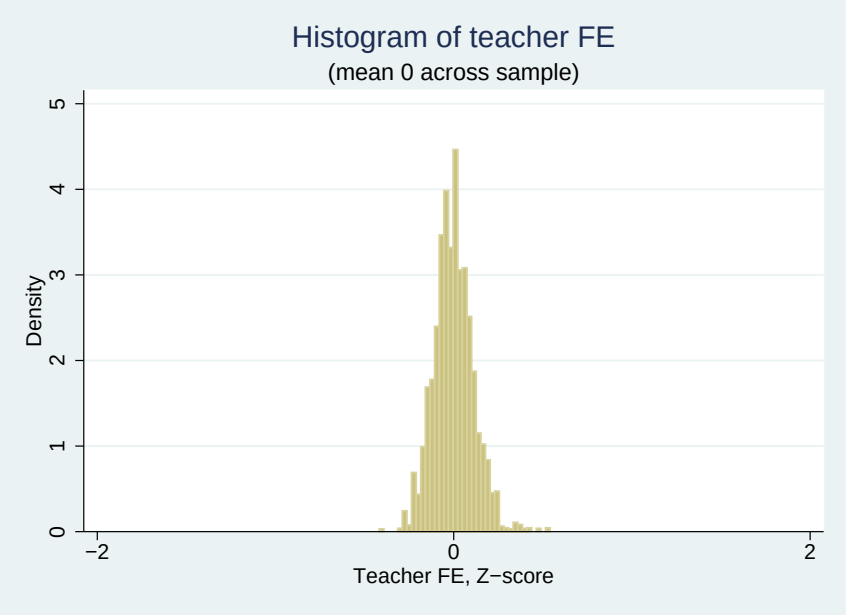

Figure 9: Teacher fixed effects, randomly assigned difference in z-scores in math, top and bottom 5% trimmed, no other controls

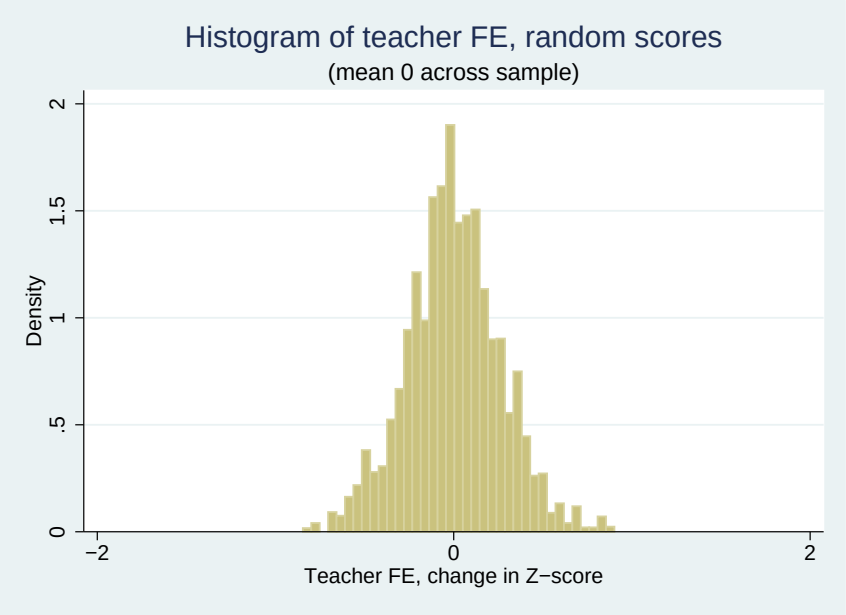

Figure 10: Teacher fixed effects, randomly assigned difference in z-scores in reading, top and bottom 5% trimmed, no other controls

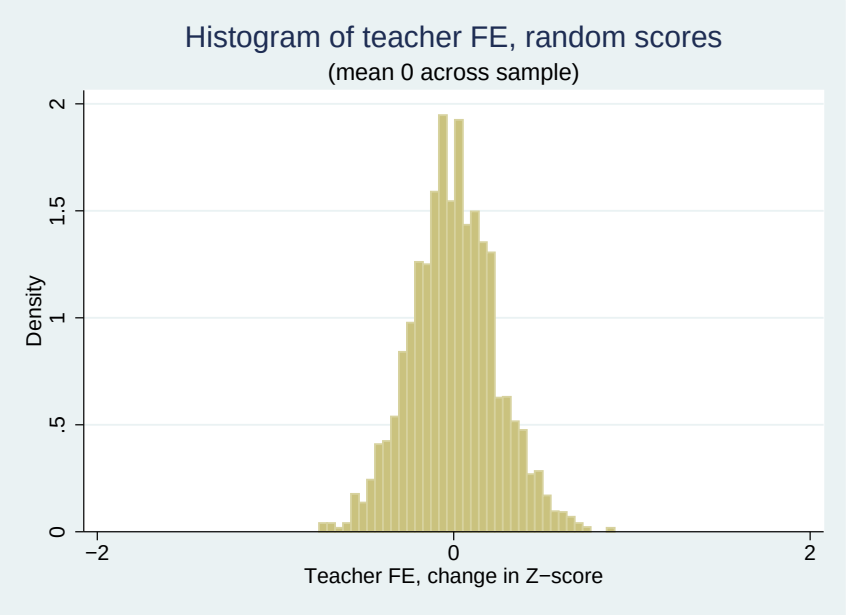

Figure 11: Teacher fixed effects, randomly assigned difference in z-scores in height, top and bottom 5% trimmed, no other controls

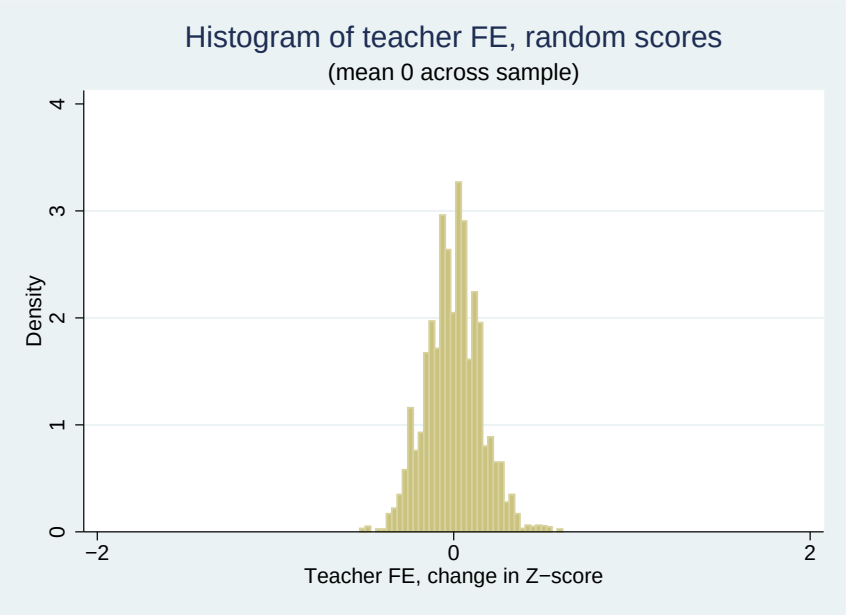

Figure 12: Teacher fixed effects, randomly assigned difference in z-scores in weight, top and bottom 5% trimmed, no other controls