Relative Skills in the Classroom:
Teachers' Gender-Differentiated Impacts on Test Scores and Course Grades

Gagandeep Sachdeva
gsachdev@ucsc.edu *
Click here for the latest version of this paper

October 23, 2025

Abstract

The gender gap in academic performance increases as students progress through school; girls outperform boys by large and increasing margins in teacher-assigned course grades and standardized reading tests, and eventually surpass boys in standardized math tests. I investigate if and how teachers affect these patterns, focusing on their gender-differentiated impacts. Using administrative data from North Carolina, I estimate value-added measures of teacher effectiveness for fifth-grade teachers, and examine their heterogeneous impacts on boys' and girls' middle school outcomes. I find that teachers with high value-added in test scores disproportionately benefit girls (particularly in math), while teachers with high value-added in course grades disproportionately benefit boys (particularly in reading). These patterns are consistent with a two-factor model in which test scores are relatively intensive in cognitive skills, course grades are relatively intensive in non-cognitive skills, and observed gender gaps imply a relative proficiency in cognitive skills for boys and a relative proficiency in non-cognitive skills for girls. Under this framework, teachers improve students most along the dimension where the students have a relative deficiency. This explanation of gender-differentiated impacts provides a unique alternative to those focused on role-model effects or teacher bias, suggesting that gender-differentiated teacher impacts reflect how teachers' strengths interact with students' underlying skill mixes.

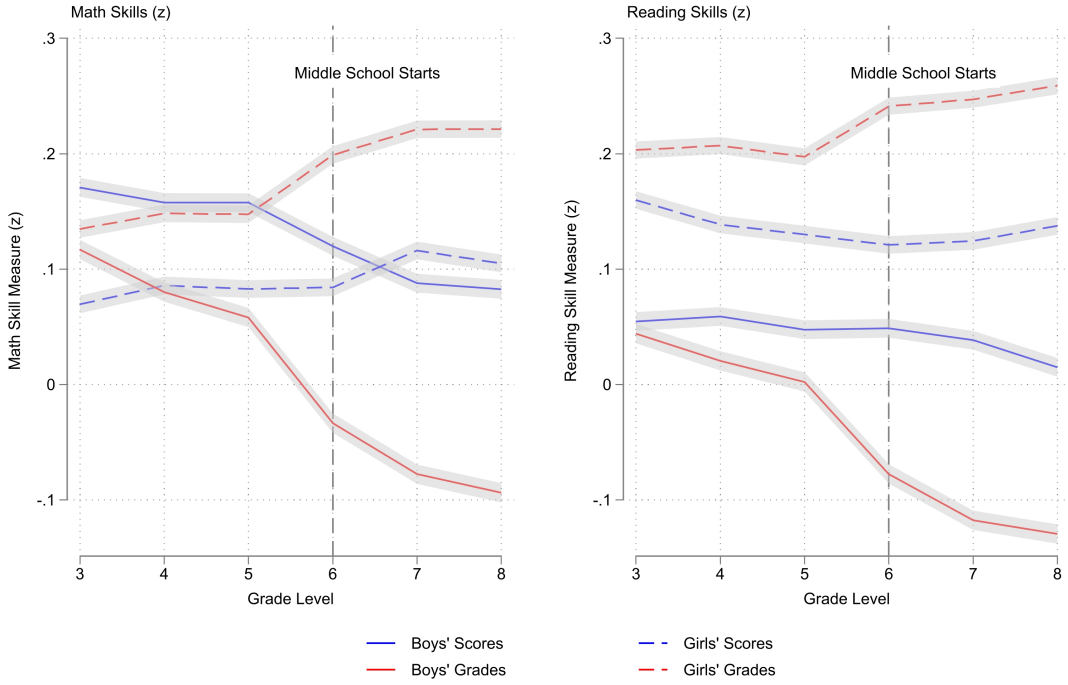
1 Introduction

Gender gaps in educational achievement vary substantially across subjects and outcome measures. In math and science, test scores often show girls lagging behind boys – especially in early grades – while teacher evaluations and course grades tend to favor girls. In reading, both test score and grade gaps tend to favor girls, with gaps in course grades being larger. These gender gaps in course grades emerge as early as elementary school, and are evident even after conditioning on test scores, suggesting that grades capture non-cognitive skills in addition to cognitive achievement (Cornwell, Mustard, and Van Parys,

*Department of Economics, UC Santa Cruz. I am grateful to Laura Giuliano, George Bulman, Robert Fairlie, and seminar participants at UC Santa Cruz for their guidance and feedback. The usual disclaimer applies.

2013; Terrier, 2020).[1] There is also evidence that achievement gaps in both test scores and grades evolve in favor of girls even more as students progress through school. Figure 1 illustrates the evolution of gender gaps between third and eighth grade for a sample of 3 cohorts in North Carolina. Test score gaps (blue lines) remain relatively stable or reverse from favoring boys to girls over time as students progress from third through eighth grade, while the female advantage in grades (red lines) grow substantially – suggesting that gender differences in non-cognitive skills (which favor girls) are increasing through elementary and middle school.

Figure 1: Math and Reading Skills: Evolution of Gender Gaps
Third–Eighth Grade



Note. Test scores and course grades are standardized within grade-year. The sample is restricted to students who can be traced from third-eighth grade. I describe more details about the sample restrictions in Section 2.1.

There is a growing body of research that speaks to the importance of gender gaps in non-cognitive measures – not least because these gaps help explain gender differences in educational attainment (Aucejo

[1]The phrase "non-cognitive skills" has been used in a variety of contexts across economics and psychology- ranging from personality traits and socio-emotional abilities to observed behaviors that predict outcomes beyond more "cognitive" measures of ability such as standardized/IQ tests (Heckman, Stixrud, and Urzua, 2006; Heckman, Pinto, and Savelyev, 2013; Lindqvist and Vestman, 2011). In the education literature, this phrase has also been applied to skills inferred from observed behaviors, such as attendance, disciplinary outcomes, or classroom engagement, that may influence teacher evaluations and/or long-term outcomes (Jackson, 2018; Petek and Pope, 2023). My use of the term is closely linked to the interpretation of Cornwell, Mustard, and Van Parys (2013), who view course grades as reflecting skills not captured by test scores (over and above the more "cognitive" skills that are reflected in standardized tests), which teachers recognize and reward even when they are not directly measured in most contexts. While the term itself is imperfect, in this paper I interpret it as capturing non-test-based dimensions of academic performance.

and James, 2019; Autor et al., 2019; Jacob, 2002).[2] There is also growing evidence that teachers have persistent impacts on both cognitive (Rivkin, Hanushek, and Kain, 2005; Chetty, Friedman, and Rockoff, 2014b) and non-cognitive skills (Jackson, 2018; Petek and Pope, 2023), and these effects are distinct from one another. Yet surprisingly, the literature has been slow to connect these two sets of findings. In particular, none has explored the questions that naturally follow: do these dual dimensions of teacher quality have systematically different impacts for boys and girls – and if so – what explains these differential effects? And how might gender gaps in a school or classroom be affected by the strengths of their teachers?

To examine these questions, I use longitudinal administrative data from North Carolina and apply a teacher value-added framework to estimate the gender-differentiated impacts of fifth-grade teachers on middle school outcomes – specifically, test scores and course grades. I start by documenting two key patterns in the data. First, girls consistently outperform boys in reading test scores between third and eighth grade by 0.08 to 0.12 standard deviations. Boys begin third grade with a 0.1 standard deviation advantage in math scores, but this advantage shrinks and reverses by seventh grade, with girls moving ahead by the end of middle school. Second, for both math and reading, girls outperform boys in course grades, with the GPA gap growing from roughly 0.1 standard deviations in third grade to about 0.35 standard deviations in eighth grade. In reading grades, the gap starts larger (about 0.16 standard deviations) and grows to roughly 0.4 standard deviations by eighth grade. These gaps persist (and indeed continue to grow) even after conditioning on test scores, as I demonstrate in Figure A.1.

To investigate the gender-differentiated effects of teachers, I estimate teacher value-added measures for fifth-grade teachers separately for each outcome, following the approach of Jackson (2018) and Petek and Pope (2023). I find that fifth-grade teachers with high value-added in course grades improve both boys' and girls' outcomes in middle school, but boys benefit significantly more.[3] A one standard deviation increase in a fifth-grade teacher's course grade value-added improves boys' middle school grades by about 0.2–0.25 standard deviations, compared to roughly 0.1 standard deviations for girls. The differential effect of about 0.15 standard deviations is statistically significant and strongest in reading grades. I observe a slightly smaller but directionally similar effect on boys' math grades, though it is not measured as precisely. These gender-differentiated effects on grades are robust across specifications, and are based on

[2]By the end of high school, boys have higher dropout rates and lower graduation rates than girls, and are also more likely to face disciplinary actions and less likely to attend college. Dropout rates in 2021: 6.1% males vs. 4.2% females (National Center for Education Statistics, 2023); on-time graduation rates in 2022-23: 84.9% vs. 89.9% (National Center for Education Statistics, 2024); suspensions/expulsions in 2020-21: boys accounted for the majority across grade levels (U.S. Department of Education, Office for Civil Rights, 2023); college enrollment in Oct. 2024: 55.4% males vs. 69.5% females among high school graduates ages 16-24 (U.S. Bureau of Labor Statistics, 2025).

[3]Value-added measures for course grades are constructed for downstream outcomes, following Jackson (2018) and Petek and Pope (2023). Specifically, course grade value-added for a fifth-grade teacher measures her average contribution to her students' course grades in sixth grade. I describe the distinction between test score and course grade value-added in greater detail in Section 3.1.

value-added measures constructed on downstream grades, thus ensuring that the results are not driven by differences in the grading practices of fifth-grade teachers.

Fifth-grade teachers with high value-added in test scores improve both boys' and girls' outcomes. In math, a one standard deviation increase in a fifth-grade teacher's test score value-added improves boys' scores by about 0.12–0.13 standard deviations, while girls' scores increase by about 0.19–0.23 standard deviations. The difference is statistically significant in later grades (specifically 7th-8th). In reading, both boys and girls benefit, with boys' scores increasing by about 0.17 standard deviations and girls' by about 0.20 standard deviations. The gender difference is smaller and not statistically significant, but the pattern is consistent with girls experiencing somewhat larger gains.

To investigate these gender-differentiated impacts further, I next construct gender-specific teacher value-added measures – separate estimates based on the boys or girls each fifth-grade teacher taught (similar to Barrios-Fernández and Riudavets-Barcons (2024) and García-Echalar, Poblete, and Rau (2024)). These measures allow me to test whether the teachers who are most effective for one gender are similarly effective for the other. I find two asymmetric patterns that reinforce the gender-differentiated impacts reported previously. Firstly, fifth-grade teachers' boy-specific test-score value-added predicts improvements in middle school test scores for both boys and girls, whereas their girl-specific test-score value-added predicts improvements only for girls. Secondly, fifth-grade teachers' girl-specific course-grade value-added predicts improvements in course grades for both boys and girls, while boy-specific course-grade value-added predicts improvements only for boys. This pattern emerges for both math and reading, and is consistent with the gender-differentiated impacts that disproportionately benefit girls in test scores and boys in grades.

Taken together, these findings are consistent with a simple theoretical framework that I develop in Section 4.1. In this framework, test scores are relatively more intensive in cognitive skills, and course grades are relatively more intensive in non-cognitive skills – a more flexible version of the assumptions made by Cornwell, Mustard, and Van Parys (2013), Jackson (2018), and Petek and Pope (2023). Under this framework, I show that (a) larger gender gaps in course grades than test scores imply a relative proficiency for boys in cognitive skills, and a relative proficiency for girls in non-cognitive skills, and (b) teachers who improve cognitive (non-cognitive) skills have stronger impacts on both test scores and grades of those students who have a relative *deficiency* in cognitive (non-cognitive) skills.[4] As a result, the

[4]Let c and n denote (unobserved) cognitive and non-cognitive skills respectively. Group A has a *relative proficiency* in cognitive skills (and a *relative deficiency* in non-cognitive skills) compared to group B if $\frac{c_A}{n_A} > \frac{c_B}{n_B}$; the reverse inequality implies a relative proficiency for Group A in non-cognitive skills and for Group B in cognitive skills. For more details, see Section 4.1.

model predicts that teachers who improve cognitive skills should disproportionately benefit girls, while teachers who improve non-cognitive skills should disproportionately benefit boys. These predictions match the patterns of my empirical findings. In Section 5.3, I provide some evidence for this interpretation by constructing a student-level measure of relative proficiency and showing that teachers have larger impacts on students in dimensions where those students have a relative deficiency.

My paper makes three contributions to the literature. First, it contributes to research on educational gender gaps and teachers' heterogeneous impacts by student gender. The literature on gender gaps in education has emphasized contexts where female students are behind, particularly in math and science test scores, track choices in middle school, and selection into STEM majors in higher education (Fryer and Levitt, 2010; Penner and Paret, 2008; Pope and Sydnor, 2010; Carlana, 2019; Delaney and Devereux, 2019; Lavy and Megalokonomou, 2024; Card and Payne, 2021). A relatively sparser literature has documented boys underperforming relative to girls in reading test scores, course grades, track placement, grade retention, suspension rates, graduation rates, and college enrollment (Fortin, Oreopoulos, and Phipps, 2015; Cornwell, Mustard, and Van Parys, 2013; Terrier, 2020; Aucejo and James, 2021; Bertrand and Pan, 2013; Jacob, 2002; Card, Chyn, and Giuliano, 2024). My paper contributes to both strands by examining outcomes across subjects where boys lag in some dimensions and girls lag in others, and by explicitly linking these gaps to teachers' differential impacts in a value-added framework. Existing work on teachers' gender-differentiated impacts has largely focused on role-model effects (Bettinger and Long, 2005; Carrell, Page, and West, 2010; Dee, 2005) or teachers' biases in evaluating students (Carlana, 2019; Lavy and Megalokonomou, 2024; Lavy and Sand, 2018; Martínez, 2025; Terrier, 2020). Only three papers explore teachers' heterogeneous impacts on boys and girls through a value-added framework, and they do so by estimating gender-specific teacher value-added measures. Aucejo et al. (2022) construct gender-specific teacher value-added in reading and find considerable heterogeneity in boy-specific and girl-specific value-added – but do not use these measures to predict outcomes for boys and girls. Barrios-Fernández and Riudavets-Barcons (2024) show that boy-specific value-added in math is higher and highlight the role of biased grading (which favors boys). García-Echalar, Poblete, and Rau (2024) find that effective teachers reduce the gender gap in math scores (which favors boys), and emphasize the role of female teachers. I show that teachers with high value-added in test scores disproportionately benefit girls, while teachers with high value-added in course grades disproportionately benefit boys. I further demonstrate that boy-specific and girl-specific value-added measures can be used to predict both boys' and girls' outcomes, with asymmetric cross-impacts of gender-specific value-added measures that reinforce these findings.

Second, I offer an interpretation of gender-differentiated TVA grounded in a simple model of relative skill deficits. Existing literature treats test scores as noisy measures of cognitive skills and course grades as noisy measures of non-cognitive skills (Cornwell, Mustard, and Van Parys, 2013; Jackson, 2018; Petek and Pope, 2023). I adopt a more flexible approach wherein test scores are relatively more intensive in cognitive skills while course grades are relatively more intensive in non-cognitive skills, and show that observed gender gaps imply a relative proficiency for boys in cognitive skills and for girls in non-cognitive skills. Under this framework, teachers who improve any given skill dimension have stronger impacts on students with a relative deficiency in that dimension. This predicts that teachers who improve cognitive skills should disproportionately benefit girls, while teachers who improve non-cognitive skills should disproportionately benefit boys – predictions that align precisely with my empirical results. This interpretation differs from explanations such as role-model effects or teachers' biases, by highlighting that the observed gender-differentiated impacts reflect baseline differences in boys' and girls' skill mixes, and how teachers' strengths interact with those mixes. Barrios-Fernández and Riudavets-Barcons (2024) and García-Echalar, Poblete, and Rau (2024) emphasize role model effects or biased grading, while Chetty, Friedman, and Rockoff (2014b) show that girls have higher long-run returns to effective teachers – but do not clarify why they see these heterogeneous impacts by gender. My framework and evidence help fill this gap.

Third, I make a methodological contribution to the literature on teachers' impacts on non-cognitive and behavioral skills. Existing literature in this space estimates teacher value-added measures using composite indices that combine grades, absences, suspensions, and grade repetition (Jackson, 2018; Petek and Pope, 2023). I decompose these measures and estimate value-added on each component separately, showing that – in the context of elementary school teachers – only course grades yield TVA measures that satisfy validation tests and persist over time. Grade repetitions are too rare to yield meaningful variation, and absences and suspensions fail standard validation tests. This suggests that course grade TVA should be treated as a standalone measure rather than combined with other behavioral measures. While grades cannot be used as an accountability tool, they remain highly predictive of student outcomes and capture teacher impacts that test scores alone miss.

The remainder of this paper is organized as follows. In Section 2, I describe the data, sample restrictions, and outline some descriptive and summary statistics for students and teachers. In Section 3, I describe the construction of teacher value-added measures, present validation and falsification tests, and introduce gender-specific value-added measures. In Section 4, I develop a theoretical framework of relative skills, show how observed empirical patterns in gender gaps align with an interpretation of relative skill deficits,

and outline the empirical specifications for estimating gender-differentiated teacher impacts. In Section 5, I present my main findings on the persistence of teacher effects and their heterogeneous impacts by student gender, and provide some evidence on how the gender-differentiated results connect to the relative skills interpretation. Finally, Section 6 concludes.

2 Data and Summary Statistics

2.1 Sample Construction

I use administrative data from the North Carolina Education Research Data Center (NCERDC), housed at Duke University in partnership with the North Carolina Department of Public Instruction. North Carolina's data has been used in several influential studies on teacher value-added, including Jacob, Lefgren, and Sims (2010), Rothstein (2010), Rothstein (2017), Jackson (2018), and Aucejo and James (2019), among several others. The NCERDC dataset covers public school students and teachers who can be tracked longitudinally, and includes detailed academic and behavioral outcomes. These features allow me to estimate teacher effects across both cognitive and non-cognitive outcomes, using complete student cohorts.

The analysis spans 2006 to 2013, based on data availability. My primary source for academic outcomes (specifically, test scores and course grades) and teacher-student linkages is the End of Grade (EOG) files in the NCERDC data. Course grades for elementary and middle school students are available from 2006 to 2013, with third grade as the first consistently observed grade. North Carolina reports "anticipated course grades" at the time of EOG testing – which are anticipated by teachers for each student's in-class performance in math and ELA, prior to finalizing report card grades. These measures are recorded before the academic year ends, and serve as proxies for final grades. While the use of anticipated rather than finalized grades introduces measurement error, Mozenter (2019) argues that this error is likely to be classical, making estimates noisier but unbiased when grades are used as outcomes. Anticipated grades are also strong predictors of high school GPA (as I show in Figure A.2), and have been used in prior work to identify variation in teacher grading patterns and its impacts on their students' outcomes in later grades. While less common compared to finalized report card grades, this system of reporting grades is still useful in the setting of North Carolina elementary schools – with self-contained classrooms, where the same teacher both instructs and evaluates students in math and reading.

In addition to the academic achievement variables, I also use three behavioral measures: number of

absences in a given grade-year, whether a student was suspended, and whether a student repeated a grade. I also construct a composite index of behavioral outcomes by combining course grades, absences, suspensions, and grade repetition, following the approach of Jackson (2018). The student characteristics I use include sex, race, economic disadvantage, English as a second language (ESL), and whether the student reported a disability.[5]

Students are linked to teachers via the EOG files, which serve as the primary source for academic outcomes. Teacher identifiers in the EOG testing files are available only through 2011, thereby establishing 2006-2013 as the estimation window for teacher value-added measures for fourth and fifth-grade teachers. Teacher value-added (VA) measures are constructed using a restricted sample that satisfies four key conditions. First, the teacher must be linked to a student in a given grade-year via the EOG file and must also be recorded as having taught a self-contained, non-special education elementary classroom for that grade-year in the School Activity Report (SAR) file. Second, VA estimates are identified only for teachers who are linked to at least 12 students in a given classroom – consistent with what is standard practice in the literature. Third, the estimation sample is limited to teachers of students in grades 4 and 5 during the years 2006 to 2011 – before which (anticipated) course grades are not available, and after which teacher identifiers are not available in the EOG files. Finally, to enable within-school estimation, schools must have a minimum of two teachers per grade in a given year. In elementary grades, the same teacher typically teaches both math and reading, enabling consistent identification of teacher effects for both subjects.

The analysis sample includes students who began third grade between 2006 and 2008, which allows me to observe the same set of students from third grade all the way through middle school (2011-2013, assuming normal progression). This design allows me to investigate the impacts of elementary teachers on middle school outcomes, while preserving the necessary student-teacher linkages for estimating value-added measures.

Table 1 reports summary statistics for key outcomes and demographics, for the full (unrestricted) sample, as well as differences between students included in and excluded from the analysis sample. Students in my analysis sample are somewhat positively selected relative to the full set of traceable students; at the end of fourth grade, they have higher test scores and course grades, fewer absences, and lower rates of grade repetition and suspension. The analysis sample also contains fewer economically disadvantaged, ESL, and disabled students, slightly smaller class sizes, and has fewer Black, Hispanic, and Asian students, though these differences are not too large. In Table A.1, I report the same statistics for third-grade students.

[5]Economically disadvantaged students are defined as those who were eligible for a free or reduced price lunch.

Table 1: Summary Statistics: Fourth-Grade Students

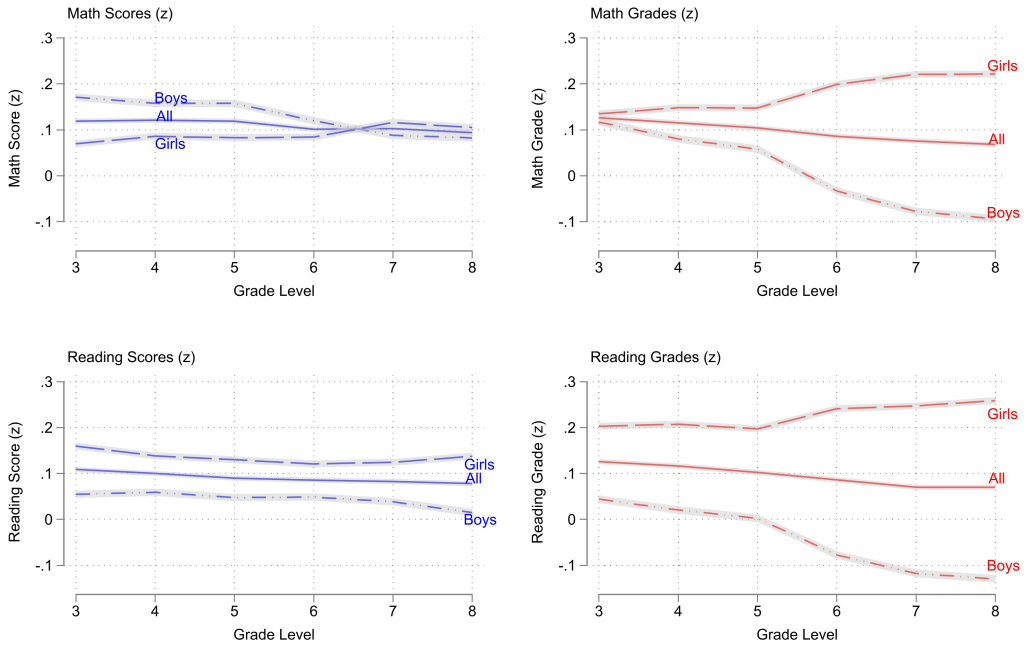
Student Outcomes			Student Characteristics		
	Mean	Diff		Mean	Diff
Test Score	0.009	0.164***	Female	0.494	0.031***
	(0.929)	(0.003)		(0.500)	(0.002)
Math Score	0.010	0.179***	White	0.549	0.055***
	(0.995)	(0.004)		(0.498)	(0.002)
Reading Score	0.002	0.158***	Black	0.266	-0.036***
	(0.996)	(0.004)		(0.442)	(0.002)
Course Grade	0.009	0.171***	Hispanic	0.106	-0.015***
	(0.940)	(0.003)		(0.308)	(0.001)
Math Grade	0.009	0.171***	Asian	0.022	-0.003***
	(0.996)	(0.004)		(0.148)	(0.001)
Reading Grade	0.008	0.174***	Other	0.057	-0.001
	(0.995)	(0.004)		(0.231)	(0.001)
Behavioral Skills	0.002	0.200***	Disadvantaged	0.492	-0.045***
	(0.998)	(0.004)		(0.500)	(0.002)
ln(1+absences)	1.672	-0.083***	ESL	0.067	-0.018***
	(0.825)	(0.003)		(0.250)	(0.001)
Suspended	0.046	-0.018***	Reported Disability	0.139	0.073***
	(0.209)	(0.001)		(0.346)	(0.001)
Repeated Grade	0.010	-0.009***	Class Size	21.559	0.406***
	(0.101)	(0.000)		(4.847)	(0.017)

Note: Reported means are for the unrestricted sample. Standard deviations are reported in parentheses below the means. Differences are computed between students included in the analysis sample and those in the unrestricted sample who are not included. Standard errors for the difference in means are reported in parentheses below the differences. All test scores and grades are standardized (z-scores). Non-cognitive skills is a standardized measure that combines course grades, absences, suspensions, and grade repetition into a single composite behavioral index, following Jackson (2018). Stars denote significance levels for a t-test for differences in means: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

2.2 Summary Statistics: Students

Table A.2 and Table 2 report summary outcomes and demographics for the 111475 students that I can trace from third through eighth grade who satisfy the sample conditions that I described earlier. Within the analysis sample, girls consistently outperform boys in reading test scores from third through eighth grade. Boys begin with a modest advantage in math test scores of about 0.10σ in third grade, but this narrows over successive grades and reverses by seventh and eighth grade, when girls pull ahead (see Figure 2). In course grades, girls outperform boys in both math and reading at every grade level. These grade gaps are substantially larger than those in test scores, and expand as students progress through middle school, increasing from 0.02σ to 0.32σ for math grades and 0.16σ to 0.39σ in reading grades between grades 3-8. As I demonstrate in Figure A.1, the gender gaps in grades (even conditioning on test scores) remain large from the outset in both subjects and grow steadily over time as well.

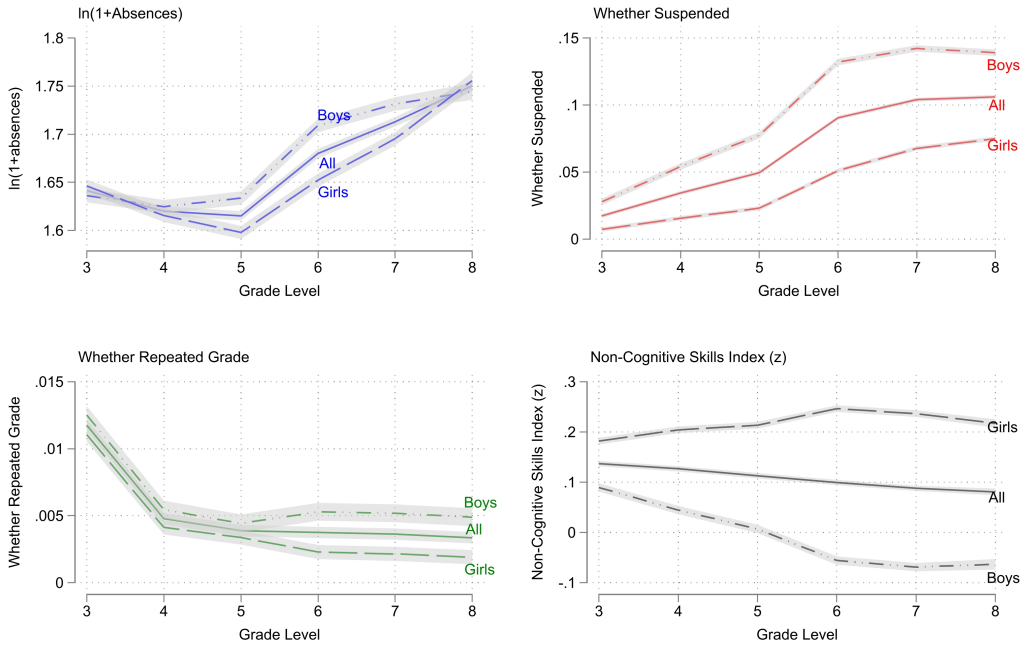
Figure 2: Academic Outcomes: Progression
Third–Eighth Grade



Note. Test scores and course grades are standardized within grade-year. The sample is restricted to students who can be traced from third-eighth grade.

Gender differences are also evident in behavioral outcomes. Boys are somewhat more likely to repeat a grade (though overall rates of grade retention are quite low in the analysis sample), consistently more likely to be suspended, and to have more absences, as I demonstrate in Figure 3.

Figure 3: Behavioral Outcomes: Progression
Third–Eighth Grade



Note. Behavioral skills index is a composite measure that combines course grades, suspensions, absences, and grade retention into a single standardized index. The sample is restricted to students who can be traced from third-eighth grade.

Demographic differences between boys and girls in the analysis sample are modest in magnitude: girls are slightly more likely to be classified as disadvantaged or disabled, while boys are marginally more likely to be White. This pattern is partly explained by differential selection into the analysis sample. Disadvantaged and Black students are less likely to be selected into the analysis sample, as I show in Table 1. Since 72 percent of Black students are classified as disadvantaged, compared to 37 percent of non-Black students, the exclusions disproportionately affect disadvantaged Black boys, leaving a relatively higher share of White boys – corroborating the findings of Autor et al. (2019). In contrast, disadvantaged girls are more likely than their male counterparts with the same characteristics to persist in the sample, which helps explain why girls in the analysis sample are somewhat more likely to be classified as disadvantaged.

2.3 Summary Statistics: Teachers

I obtain the data for teachers' observable characteristics from the personnel files in the School Activity Reports. Table A.3 reports summary statistics for teachers' gender, race, education and experience for the full set of eligible teachers, and teachers in the analysis sample of students. The analysis sample of teachers closely resembles the full set in terms of observable characteristics. Roughly 87-88 percent of

Table 2: Summary Statistics: Student Demographics (Analysis Sample)

	Mean	Diff		Mean	Diff
Disadvantaged	0.46	-0.02***	White	0.58	0.02***
	(0.50)	(0.00)		(0.49)	(0.00)
ESL	0.06	0.01***	Black	0.24	-0.02***
	(0.23)	(0.00)		(0.43)	(0.00)
Reported Disability	0.09	0.06***	Hispanic	0.10	0.00
	(0.28)	(0.00)		(0.30)	(0.07)
Class Size (3rd)	19.76	0.03	Asian	0.02	-0.00
	(3.09)	(0.15)		(0.14)	(0.32)
Class Size (4th)	21.81	0.05*	Other	0.06	-0.00**
	(3.44)	(0.02)		(0.23)	(0.00)
N	111475			111475	

Note: Reported means are for the analysis sample. Standard deviations are reported in parentheses below the means. Differences are computed between boys and girls in the analysis sample. Standard errors for the difference in means are reported in parentheses below the differences. Stars denote significance levels for a t-test for differences in means: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

teachers are female. About 85 percent are White, 14 percent are Black, with very few Hispanic teachers. Around one-third of teachers hold a graduate degree. Teachers in the analysis sample are also slightly more experienced on average (10.1 years versus 9.1 years).

3 Teacher Value-Added Measures

For estimating teacher value-added measures, I follow an approach adapted from Kane and Staiger (2008), Jackson (2018), and Petek and Pope (2023), with refinements from more recent literature that I describe below.

3.1 Estimating Teacher Value-Added Measures

I begin by estimating teacher value-added with test scores as outcomes. For a student i in classroom c, school s, and cohort t, I model the grade-g test score as

$$y^g_{icst} = \beta_0 + \mathbf{Y}^{g-1}_{icst}\boldsymbol{\beta}_1 + \mathbf{X}_i\boldsymbol{\beta}_2 + \gamma^{g-1}_{cst} + \phi^g_s + \epsilon^g_{icst} \tag{1}$$

Here, y^g_{icst} is the test score of student i in grade g, $\mathbf{Y}^{g-1}_{icst}$ is the vector of lagged individual, classroom, and school outcomes from grade $g-1$, $\mathbf{X}_i$ is a vector of student demographics, γ^{g-1}_{cst} denotes fixed effects for student i's classroom in grade $g-1$ (thereby accounting for the role of prior teachers in any potential

sorting, as described by Gilraine and McCarthy (2024) and Staiger, Kane, and Johnson (2024)), and ϕ_s^g denotes fixed effects for student i's school in grade g.[6] Specifically, $\mathbf{Y}_{icst}^{g-1}$ includes subject-specific test scores and course grades, as well as absences, and flags for grade repetition and suspensions for student i in grade $g-1$, following the multidimensional value-added framework laid out by Jackson (2018). $\mathbf{Y}_{icst}^{g-1}$ also includes leave-one-out classroom and school-level averages of course grades, test scores, absences, suspensions, and grade retention – to account for the potential to be assigned to a given teacher based on prior-grade peer characteristics (Clotfelter, Ladd, and Vigdor, 2006; Horváth, 2015). All components of $\mathbf{Y}_{icst}^{g-1}$ enter the model as cubic polynomials to account for non-linearities in how these outcomes grow, following standard practice in the literature. The vector $\mathbf{X}_i$ includes student i's race, gender, whether they had english as a second language, whether they were economically disadvantaged, and whether they reported a disability. The residual ϵ_{icst}^g is the component of achievement not predicted by prior outcomes, demographics, or peer composition.

For estimating value-added measures for outcomes other than test scores, I modify this setup by using the lead of the outcome (i.e., y_{icst}^{g+1}), mirroring the approach followed by Jackson (2018) and Petek and Pope (2023). While this introduces more noise in the model (since it partially captures the effect of the teacher in grade $g+1$ as well), it reduces any potential biases stemming from grading practices differing between teachers, or from teachers evaluating their own students. Specifically, sixth-grade course grades are regressed on fourth-grade outcomes and covariates to form value-added measures for fifth-grade teachers. The right-hand-side controls mirror those in equation (1), with lagged subject-specific test scores and course grades included in $\mathbf{Y}_{icst}^{g-1}$.

For behavioral outcomes such as absences, suspensions, and grade repetition, I also use leads as the dependent variable to estimate fifth-grade teacher effects. In this case, the vector of prior outcomes $\mathbf{Y}_{icst}^{g-1}$ includes test scores and grades averaged across subjects, together with lagged behavioral measures. This ensures that teacher value-added in behavioral outcomes is identified from the persistence of their effects into subsequent grades, while conditioning flexibly on students' prior academic and behavioral outcomes.

Following Kane and Staiger (2008), Koedel, Mihaly, and Rockoff (2015), and Petek and Pope (2023), the residual ϵ_{icst}^g can be additively decomposed into a teacher value-added term δ_j^g, a classroom-specific term μ_c^g, and a student-specific idiosyncratic error term ν_{ist}^g.[7]

$$\epsilon_{icst}^g = \delta_j^g + \mu_c^g + \nu_{icst}^g \tag{2}$$

[6] I define a classroom in elementary school as a school-grade-teacher-cohort tuple.

[7] Chetty, Friedman, and Rockoff (2014a) and other literature allow for growth in a teacher's value-added measure with experience by including a drift term in the model. I refrain from this, since my estimations are based on a fewer number of eligible cohorts.

I estimate leave-cohort-out versions of Equation (1), and extract the residuals $\widehat{\epsilon}_{icst}$ from these regressions. I then average the residuals for each teacher j:

$$\widehat{\delta}_j^g = \frac{1}{n_j^g} \sum_{i \in j} \widehat{\epsilon}_{icst}^g \tag{3}$$

$\widehat{\delta}_j^g$ provides an unbiased estimate of δ_j^g as long as $\mathbb{E}[\mu_c^g + \nu_{icst}^g | j] = \mathbb{E}[\mu_c^g + \nu_{icst}^g]$, i.e., conditional on the controls specified in Equation (1), teachers don't receive students that systematically differ on unobserved achievement. Chetty, Friedman, and Rockoff (2014a) and Bacher-Hicks, Kane, and Staiger (2014) show that the controls used for students' prior achievement and behavior account for most student sorting in value-added models, and Gilraine and McCarthy (2024) and Staiger, Kane, and Johnson (2024) show that the inclusion of prior teacher fixed effects (captured through γ_{cst}^{g-1} in Equation (1)) further absorbs unobserved variation in student achievement that could otherwise generate bias through sorting. To alleviate concerns about whether students differentially sort into fifth-grade classrooms taught by better teachers, I perform a placebo test wherein I estimate the effects of high value-added fifth-grade teachers on fourth-grade outcomes, following Rothstein (2010) and Rothstein (2017). I show in Figure 5 that conditional on a third-grade baseline, fifth-grade teacher value-added does not predict fourth-grade outcomes for test scores and course grades. Finally, I use empirical bayesian methods to shrink the teacher effects towards the mean of the distribution, adapting the approach of Chetty, Friedman, and Rockoff (2014a) and Angrist et al. (2017), as shown below.

$$\widehat{\delta}_j^{EB} = \underbrace{\left(\frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{s}_j^2} \right)}_{1-a_j} \widehat{\delta}_j + \underbrace{\left(\frac{\hat{s}_j^2}{\hat{\sigma}_\delta^2 + \hat{s}_j^2} \right)}_{a_j} \bar{\delta} \tag{4}$$

Here, a_j is the shrinkage weight associated with teacher j; higher values of a_j imply a greater amount of shrinkage towards the common mean (i.e., $\bar{\delta}$) for teacher j, due to noisier estimates of $\widehat{\delta}_j$. s_j is the standard error of $\widehat{\delta}_j$, and $\hat{\sigma}_\delta^2 = \frac{1}{J} \sum_{j=1}^{J} [(\widehat{\delta}_j - \bar{\delta})^2 - \hat{s_j}^2]$. $\hat{s_j}^2$ is subtracted as a bias-correction term that accounts for the excess variance in $\widehat{\delta}_j$ due to sampling error (Angrist et al., 2017).

3.2 Descriptive and Summary Statistics: Value-Added Measures

I construct value-added measures for several academic and behavioral outcomes. For test scores, I estimate teacher value-added in math, reading, and the average across subjects. For course grades, I similarly

estimate teacher value-added in math, reading, and the average across subjects. Beyond these measures, I estimate value-added for three behavioral outcomes: grade retention, suspensions, and absences. Finally, I construct value-added measures for a composite index that aggregates course grades, retention, suspensions, and absences, following the approach of Jackson (2018).

The diagonal elements of Table 3 report the standard deviations of the constructed value-added measures, expressed in standard deviations of student outcomes.[8] Test score value-added has standard deviations of 0.104 overall, 0.135 for math, and 0.115 for reading – magnitudes that fall within the range documented in the literature on teacher effects on test scores (e.g., Hanushek and Rivkin, 2010; Koedel, Mihaly, and Rockoff, 2015). Course grade value-added has standard deviations of 0.132 overall, 0.153 for math, and 0.154 for reading. The value-added measures for behavioral outcomes are somewhat more dispersed; the standard deviations are 0.160 for absences, 0.181 for suspensions, and 0.192 for grade retention.

The off-diagonal elements show correlations across value-added measures. Math and reading value-added are highly correlated within assessment type; the correlation coefficients between math test scores and reading test scores is 0.407 (similar to the estimates obtained by Goldhaber, Cowan, and Walch (2013)), and the correlation coefficients between math and reading course grades is 0.449. Correlations across assessment types are smaller: test score and course grade value-added have a correlation coefficient of 0.186 overall (Petek and Pope (2023) report similar correlation coefficients), with the coefficient between math scores and math grades being higher than that between reading scores and reading grades (0.212 vs 0.149). Value-added measures for behavioral outcomes correlate weakly with test score value-added, with correlation coefficients of 0.049 for absences, 0.007 for suspensions, and 0 for grade repetition. By contrast, they correlate more strongly with course grade value-added, with correlation coefficients of 0.130, 0.115, and 0.064, respectively – which is again consistent with the patterns reported by Petek and Pope (2023).

[8] I report the standard deviations of the unshrunk fixed effects- in order to facilitate comparisons with standard deviations reported in other literature.

Table 3: Correlations and Standard Deviations of Fifth-Grade Teacher Value-Added Measures

	Test Score	Math Score	Reading Score	Course Grade	Math Grade	Reading Grade	Behavioral Skills	Absences	Suspensions	Grade Repetition
Test Scores	0.104									
Math Scores	0.864	0.135								
Reading Scores	0.784	0.407	0.115							
Course Grades	0.186	0.177	0.138	0.132						
Math Grades	0.185	0.212	0.102	0.837	0.153					
Reading Grades	0.143	0.102	0.149	0.832	0.449	0.154				
Behavioral Skills	0.121	0.115	0.086	0.589	0.500	0.504	0.159			
Absences	0.049	0.049	0.038	0.130	0.123	0.108	0.511	0.160		
Suspensions	0.007	0.004	-0.000	0.115	0.089	0.106	0.631	0.127	0.181	
Grade Repetition	0.000	-0.009	0.003	0.064	0.049	0.055	0.236	0.030	0.077	0.192

Note: The table reports correlations (below the diagonal) and standard deviations (on the diagonal) of teacher value-added measures for fifth-grade teachers.

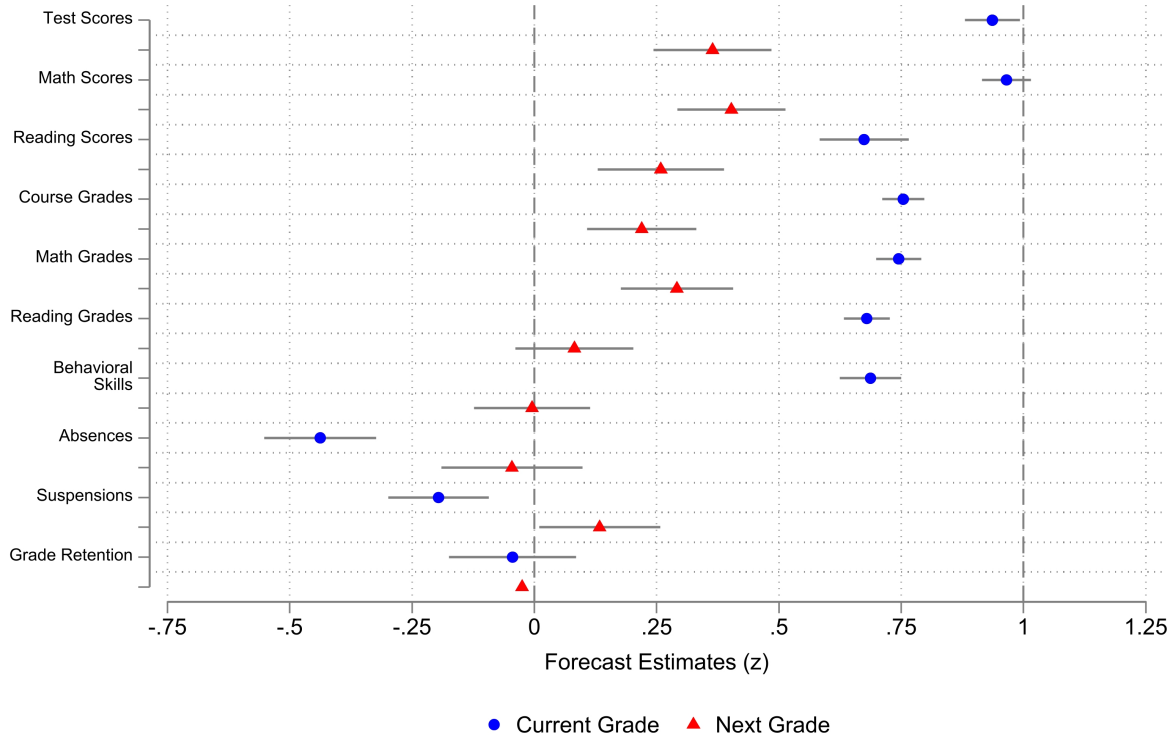
3.3 Validity and Falsification Tests

To assess the validity of the estimated teacher value-added measures, I follow the approach developed by Chetty, Friedman, and Rockoff (2014a) and test whether these measures are forecast-unbiased. Intuitively, a teacher who improve student achievement by one standard deviation should, on average, generate an improvement of one standard deviation in student outcomes. Formally, I estimate the following equation:

$$y_{icst}^{h} = \alpha_0 + \alpha_1 \widehat{\delta}_j^g + \mathbf{Y}_{icst}^{g-1}\alpha_3 + \mathbf{X}_i\alpha_4 + \gamma_{cst}^{g-1} + \phi_s^g + u_{icst}^{h} \tag{5}$$

where y_{icst}^{h} is student i's outcome in grade h, $\widehat{\delta}_j^g$ is the leave-cohort out empirical Bayes estimate of the value-added measure for student i's grade-g teacher for outcome y, and all other variables are as defined in Section 3.1. I test forecast-unbiasedness for value-added measures based on both contemporaneous (i.e., $h = g$) and lead ($h = g + 1$) values of y for all outcomes. Following Chetty, Friedman, and Rockoff (2014a), the VA measures are defined as *forecast-unbiased* if $\hat{\alpha_1} = 1$. I report my estimates of $\hat{\alpha_1}$ for all value-added measures for fifth-grade teachers in Figure 4.

Figure 4: Validations of Value-Added Measures
Fifth-Grade Teachers



Note. Figure reports estimated values of $\hat{\alpha_1}$ for value-added measures constructed for test scores, course grades, and behavioral outcomes, measured in standard deviation units of the outcomes. Estimated coefficients are from regressions wherein fifth (current) and sixth (next) grade outcomes are regressed on fifth-grade VA measures.

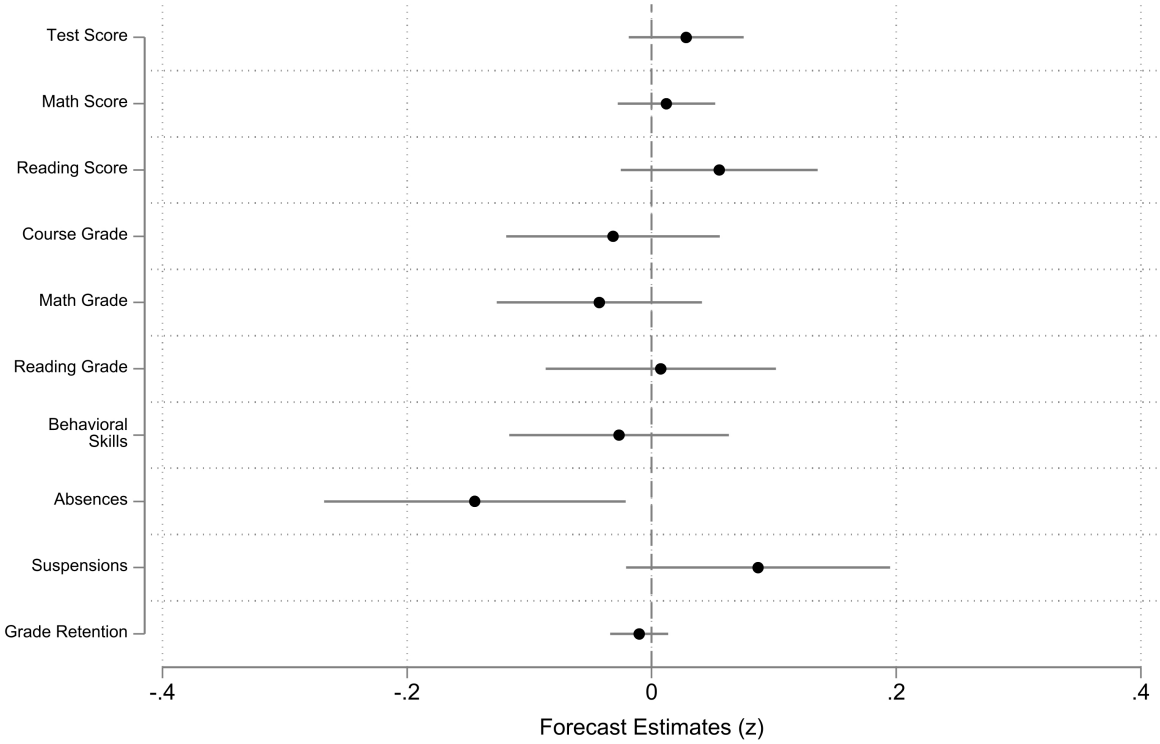
The coefficients are closest to one for test score VA measures, but smaller for course grade VA and smaller still for behavioral VA measures. This pattern is consistent with two facts. First, value-added measures constructed for lead outcomes (as in the case of grades and behavioral measures) necessarily introduce more noise, since they partially reflect the effects of the teacher and classroom environment of grade $g+1$, in addition to those in grade g. Second, test scores are inherently less noisy than course grades, since test scores are continuous and based on a given set of standards for all students within a given grade-year, and course grades are discrete and teacher-assigned. This is precisely what I find for value-added measures in my analysis. I show in Figure B.2 that (a) the standard errors for the (unshrunk) value-added measures defined on lead outcomes are always higher than the standard errors for the value-added measures on contemporaneous outcomes, and (b) the standard errors for the value-added measures on course grades and behavioral outcomes are higher than the standard errors for the value-added measures for test scores.[9] Taken together, these facts imply that one should expect the estimates of $\hat{\alpha_1}$ to be more attenuated for lead outcomes relative to contemporaneous outcomes, and more attenuated for non-test score measures. This is precisely what I show in Figure 4. The results therefore suggest that the positive and significant coefficients for course grade VA should be interpreted as downward-biased estimates of teachers' true impacts. However, for absences, suspensions, and grade repetition, I cannot conclusively reject that $\hat{\alpha_1} = 0$. Furthermore, I show in Table B.2 that – in addition to not predicting their respective outcomes in sixth grade – value-added measures for absences, suspensions, and grade repetition do not consistently predict future absences, suspensions, or grade repetition throughout middle school. It is also worth pointing out here that even in the unrestricted sample, only 1% of students repeated fourth grade, and the number is even smaller for the positively selected analysis sample (as I show in Table 1). Grade retention in elementary school is a relatively rare occurrence, and therefore there may not be enough variation in the outcome to yield meaningful teacher value-added measures for grade repetition.

Furthermore, I implement the falsification design proposed by Rothstein (2010) and Rothstein (2017)- regressing students' fourth-grade outcomes on the value-added measure of their fifth-grade teachers. I use the same specification as outlined in Equation (5) – with two key differences: $h = g - 1$, and the lagged outcomes and classroom fixed effects are defined for third grade (i.e., $g - 2$). If the value-added measures are seen as credible measures of a teacher's effectiveness, future teachers should not influence past outcomes, and the estimate of $\hat{\alpha}_1$ should be close zero. This is precisely what I find for test score and course grade VA measures – as I show in Figure 5.[10]

[9] I use the standard errors of the unshrunk value-added measures since they are a better indicator of how noisy the measures are. Higher standard errors also imply a greater degree of shrinkage- see Equation (4).

[10] Test score value-added is defined using fifth-grade test scores, and all other value-added measures are defined using sixth-grade outcomes.

Figure 5: Falsification Tests for Value-Added Measures
Fifth-Grade Teachers



Note. Figure reports estimated values of $\hat{\alpha_1}$ for value-added measures constructed for test scores, course grades, and behavioral outcomes, measured in standard deviation units of the outcomes. Estimated coefficients are from regressions wherein 4th grade outcomes are regressed on 5th grade VA measures.

However, value-added measures for absences and suspensions for fifth-grade teachers are weakly predictive of absences and suspensions in fourth grade, thereby questioning the validity of these measures in the context of my data. Taken together, these results reinforce two key conclusions. First, test score VA passes both validation and falsification tests cleanly, and provides the most precise measure of teacher effectiveness. Second, among the non-test-score outcomes, only course grade VA yields measures that are credible. The other behavioral components (absences, suspensions, and grade repetition) either fail the standard validation checks, or lack sufficient variation to produce reliable estimates. While the composite behavioral skills index performs reasonably well across the two tests, its validity rests almost entirely on the contribution of course grades. This motivates treating course grade VA as a stand-alone measure, rather than using it as a component of the behavioral skills index. For these reasons, I focus on test score VA and course grade VA as the two focal measures of teacher effectiveness for the remainder of this paper.

3.4 Gender-Specific Value-Added Measures

To further explore teachers' gender-differentiated impacts, I construct boy-specific and girl-specific value-added measures. The procedure mirrors what I describe in Section 3.1, but I restrict the sample to boys for boy-specific measures and to girls for girl-specific measures. To ensure that the measures are based on a sufficient amount of variation, I limit the sample to classrooms that contain at least six boys and six girls in a given grade-year, rather than the minimum of twelve students used for the standard measures. The gender-specific measures provide a way to test whether a teacher's effectiveness for one gender predicts outcomes for the other, and to disentangle whether teachers' impacts on gender gaps arise from differential impacts for boys, for girls, or for both. I report some descriptive relationships between gender-specific value-added measures in Section B.5.

4 Gender-Differentiated Impacts of Teachers

4.1 Theoretical Framework

I model student achievement as a function of two latent skills: cognitive skills (c) and non-cognitive skills (n). These skills combine as two imperfectly substitutable inputs to produce two observable outcomes: standardized test scores (y) and course grades (z). I operationalize this using a simple Cobb-Douglas function:

$$y = c^{\theta} n^{1-\theta}, \qquad \text{and} \qquad z = c^{\gamma} n^{1-\gamma}$$

Prior literature has typically treated test scores as cognitive skill measures, and course grades as non-cognitive skill measures (Cornwell, Mustard, and Van Parys, 2013; Jackson, 2018; Petek and Pope, 2023). I adopt a more flexible approach and assume that test scores are relatively more intensive in cognitive skills, and course grades are relatively more intensive in non-cognitive skills. In the context of the expressions above, this implies that $0 < \gamma < \theta < 1$.

Let students' gender be indexed by $j \in \{b, g\}$ (boys, girls) and subjects by $s \in \{m, r\}$ (math, reading). I define observed relative outcomes for subject s between boys and girls as

$$\eta_s^y = \frac{y_{bs}}{y_{gs}}, \qquad \text{and} \qquad \eta_s^z = \frac{z_{bs}}{z_{gs}}$$

and I define the ratio of latent skills as

$$\psi_{js} = \frac{c_{js}}{n_{js}}$$

It follows that

$$\frac{\psi_{bs}}{\psi_{gs}} = \left(\frac{\eta_s^y}{\eta_s^z}\right)^{\frac{1}{\theta-\gamma}} \tag{6}$$

This expression links observed gender differences in test scores and grades to differences in the underlying relative mixes of unobserved cognitive and non-cognitive skills.

4.1.1 Prediction

Two stylized facts in my data guide the interpretation of gender differences in test scores and grades. In mathematics, boys begin with higher test scores but lower grades than girls, so that $\eta_m^y > 1$ while $\eta_m^z \leq 1$. This implies $\eta_m^y/\eta_m^z > 1$, and therefore boys have a higher ratio of cognitive to non-cognitive skills than girls (i.e., $\psi_{bm} > \psi_{gm}$). In reading, girls outperform boys in both test scores and grades, but the grade gap is larger than the test score gap. Thus, $\eta_r^y/\eta_r^z > 1$, which again implies that boys have a higher ratio of cognitive to non-cognitive skills than girls (i.e., $\psi_{br} > \psi_{gr}$). Taken together, these stylized facts imply that $\psi_{bs} > \psi_{gs}$ for both subjects, and the differences in relative skills for boys and girls can be interpreted as boys having a relative proficiency in cognitive skills, and girls having a relative proficiency in non-cognitive skills for both subjects. It is important to note here that this interpretation of relative proficiency does not make any assumptions about which gender has greater (or lower) absolute levels of cognitive or non-cognitive skills.

From this starting point, the gender-differentiated effects of teachers can be explained as follows. Teachers disproportionately improve students in the dimension where they have a relative *deficiency*. A teacher who improves cognitive skills (c) will therefore improve girls' outcomes more than boys', holding fixed her influence on non-cognitive skills (n). Specifically, when a teacher improves cognitive skills, the ratio of girls' improvements to boys' improvements is given by $\left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{1-\theta}$ for test scores and $\left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{1-\gamma}$, both of which are greater than 1. Conversely, a teacher who improves non-cognitive skills (n) will improve boys' outcomes more than girls', holding fixed her influence on cognitive skills. Specifically, when a teacher improves non-cognitive skills, the ratio of boys' improvements to girls' improvements is given by $\left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{\theta}$ for test scores and $\left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{\gamma}$ for grades, both of which are greater than 1.[11]

These predictions apply across both math and reading, even though the direction of raw gaps differs. For

[11] I formally derive these expressions in Section C

math scores and grades, girls' relative deficiency in cognitive skills means they benefit more from teachers who improve test scores, while boys' relative deficiency in non-cognitive skills means they benefit more from teachers who improve grades. The same logic applies for reading outcomes: the larger grade gap points to boys' relative deficiency in non-cognitive skills, so teachers who improve grades disproportionately benefit boys, while teachers who improve test scores disproportionately benefit girls.

4.2 Empirical Design

The predictions from Section 4.1 imply that teachers who improve cognitive skills should disproportionately benefit girls, while teachers who improve non-cognitive skills should disproportionately benefit boys. I evaluate these predictions in two ways.

4.2.1 Baseline Specification

To test for heterogeneous impacts of teachers on boys and girls, I begin by interacting teacher value-added measures with an indicator for whether the student is female. I estimate a multidimensional version (similar to Jackson (2018)) of Equation (5), as shown below:

$$y_{icst}^{h} = \alpha_0 + \alpha_1 F_i + \boldsymbol{\delta}_j^g \alpha_2 + \left(\boldsymbol{\delta}_j^g \times F_i\right) \alpha_3 + \mathbf{Y}_{icst}^{g-1} \alpha_4 + \mathbf{X}_i \alpha_5 + \gamma_{cst}^{g-1} + \phi_s^g + u_{icst}^{h} \tag{7}$$

Here F_i is an indicator for whether student i is female, $\boldsymbol{\delta}_j^g = \left[\widehat{\delta}_j^{g,\,\text{score}} \;\; \widehat{\delta}_j^{g,\,\text{grade}}\right]$ is the vector of teacher value-added in test scores and course grades for teacher j in grade g, and $\mathbf{Y}_{icst}^{g-1}$ and $\mathbf{X}_i$ are the controls as specified in Equation (5), except they are also interacted with student gender. The coefficient vectors $\alpha_2 = \left[\alpha_2^{\text{score}}, \; \alpha_2^{\text{grade}}\right]$ and $\alpha_3 = \left[\alpha_3^{\text{score}}, \; \alpha_3^{\text{grade}}\right]$ separately capture the impacts of score and grade value-added. α_2 reflects the effects for boys (the baseline group), while α_3 captures the differential effects for girls relative to boys. A test for the predictions outlined in Section 4.1 is α_3^{score} should be positive (i.e., test score value-added – which serves as a measure of a teacher's contribution to cognitive skills – should have stronger impacts on girls), and α_3^{grade} should be negative (i.e., course grade value-added – which serves as a measure of a teacher's contribution to non-cognitive skills – should have stronger impacts on boys).

4.2.2 Gender-Specific Value-Added

Next, I investigate whether teachers' boy-specific and girl-specific value-added measures have heterogeneous impacts on boys' and girls' future outcomes. To do this, I adapt the specification used by Barrios-Fernández and Riudavets-Barcons (2024), and estimate the following model:

$$\begin{aligned} y^h_{icst} = \alpha_0 + \left(\boldsymbol{\delta}^{g,B}_j \times M_i\right)\alpha_1 + \left(\boldsymbol{\delta}^{g,G}_j \times M_i\right)\alpha_2 + \left(\boldsymbol{\delta}^{g,B}_j \times F_i\right)\alpha_3 + \left(\boldsymbol{\delta}^{g,G}_j \times F_i\right)\alpha_4 \\ + \mathbf{Y}^{g-1}_{icst}\alpha_6 + \mathbf{X}_i\alpha_7 + \gamma^{g-1}_{cst} + \phi^g_s + u^h_{icst} \end{aligned} \tag{8}$$

Here F_i is an indicator for whether student i is female, and $M_i = 1 - F_i$ is an indicator for whether student i is male. $\boldsymbol{\delta}^{g,B}_j = \left[\widehat{\delta}^{g,B,\,\text{score}}_j,\ \widehat{\delta}^{g,B,\,\text{grade}}_j\right]$ denotes the vector of boy-specific teacher value-added for teacher j in grade g, and $\boldsymbol{\delta}^{g,G}_j = \left[\widehat{\delta}^{g,G,\,\text{score}}_j,\ \widehat{\delta}^{g,G,\,\text{grade}}_j\right]$ denotes the vector of girl-specific teacher value-added. The controls $\mathbf{Y}^{g-1}_{icst}$ and $\mathbf{X}_i$ are the lagged outcomes and student demographics defined in Equation (7).

The coefficient vectors $\alpha_1 = \left[\alpha_1^{\text{score}},\ \alpha_1^{\text{grade}}\right]$, $\alpha_2 = \left[\alpha_2^{\text{score}},\ \alpha_2^{\text{grade}}\right]$, $\alpha_3 = \left[\alpha_3^{\text{score}},\ \alpha_3^{\text{grade}}\right]$, and $\alpha_4 = \left[\alpha_4^{\text{score}},\ \alpha_4^{\text{grade}}\right]$ capture the impacts of boy-specific and girl-specific teacher value-added on both boys and girls. Specifically, α_1 reflects the effect of boy-specific VA on boys, α_2 reflects the effect of girl-specific VA on boys, α_3 reflects the effect of boy-specific VA on girls, and α_4 reflects the effect of girl-specific VA on girls.

This equation estimates four sets of coefficients for each outcome (test scores and grades), corresponding to the four student-teacher interactions. Comparing these coefficients, one can directly test the predictions outlined in Section 4.1. Specifically, if teachers who improve cognitive skills disproportionately benefit girls, then one should observe $\alpha_3^{\text{score}} + \alpha_4^{\text{score}} > \alpha_1^{\text{score}} + \alpha_2^{\text{score}}$. Similarly, if teachers who improve non-cognitive skills disproportionately benefit boys, one should observe $\alpha_1^{\text{grade}} + \alpha_2^{\text{grade}} > \alpha_3^{\text{grade}} + \alpha_4^{\text{grade}}$.

5 Results

5.1 Persistence of Teacher Effects

I begin by examining whether the estimated teacher value-added measures capture meaningful and persistent impacts on student achievement, before turning to their gender-differentiated impacts. Formally, I estimate a version of Equation (7) that is not differentiated by student gender, and allows for both dimen-

sions of a teacher's effectiveness (captured through test score value-added and course-grade value-added) to influence any given outcome, following Jackson (2018).

$$y^h_{icst} = \alpha_0 + \boldsymbol{\delta}^g_j \alpha_1 + \mathbf{Y}^{g-1}_{icst} \alpha_2 + \mathbf{X}_i \alpha_3 + \gamma^{g-1}_{cst} + \phi^g_s + \zeta_h + u^h_{icst} \tag{9}$$

Here, y^h_{icst} is an outcome (test score or course grade) for student i in grade $h \geq g+1$, and $\boldsymbol{\delta}^g_j = \left[\widehat{\delta}^{g,\text{ score}}_j \quad \widehat{\delta}^{g,\text{ grade}}_j\right]$ is the vector of test score and course grade value-added measures for teacher j who taught student i in grade g. The vectors $\mathbf{Y}^{g-1}_{icst}$ and $\mathbf{X}_i$, and the fixed effects γ^{g-1}_{cst} and ϕ^g_s are as specified for Equation (1). ζ_h denotes fixed effects for the grade level in which the outcome is measured. Both value-added measures are leave-cohort-out estimates of teacher effects that are shrunk using the empirical Bayesian approach described in Section 3.1. Test score value-added for a fifth-grade teacher is constructed using fifth-grade test scores, and course grade value-added for a fifth-grade teacher is constructed using sixth-grade course grades. Standard errors in all models are clustered at the level of the fifth-grade teacher.

Table 4: Multidimensional Impacts of Fifth-Grade Teachers on Middle School Outcomes

	Test Scores				Course Grades			
	6th-8th		7th-8th		6th-8th		7th-8th	
	I	II	III	IV	V	VI	VII	VIII
Test Score VA	0.190***	0.191***	0.147***	0.148***		0.088***		0.036
	(0.022)	(0.022)	(0.024)	(0.025)		(0.031)		(0.036)
Course Grade VA		-0.007		-0.006	0.195***	0.170***	0.182***	0.172***
		(0.031)		(0.034)	(0.044)	(0.045)	(0.049)	(0.051)
Constant	-0.142	-0.141	-0.055	-0.055	-0.104	-0.102	-0.020	-0.020
	(0.182)	(0.182)	(0.205)	(0.205)	(0.182)	(0.182)	(0.218)	(0.218)
Mean	0.091	0.091	0.089	0.089	0.076	0.076	0.071	0.071
N	334425	334425	222950	222950	334425	334425	222950	222950
r^2	0.769	0.769	0.758	0.758	0.463	0.463	0.450	0.450

Note: Standard errors clustered at the fifth-grade teacher level. Test Score VA is defined on fifth-grade test scores. Course Grade VA is defined on sixth-grade course grades. All outcomes are standardized (z-scores). Models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4 reports the impacts of fifth-grade teacher value-added on middle school outcomes, averaged over subjects. Both test score and course grade value-added measures are significant and persistent predictors of test scores and course grades respectively. A 1σ increase in teacher test score value-added improves students' middle school test scores by about 0.19σ on average (column I). This effect is robust to the inclusion of the same teacher's value-added in course grades (column II). Similarly, a 1σ increase in course grade value-added improves students' middle school grades by about 0.19σ (column V). Controlling for

test score value-added reduces the estimate slightly to 0.17σ (column VI), but the effect remains large and statistically significant.

Since course grade value-added is constructed using sixth-grade course grades, it is possible that a part of the effect for outcomes in sixth-eighth grades reflects the mechanical persistence of the construct for sixth-grade outcomes. To address this, I separately estimate impacts on outcomes measured in seventh and eighth grades. Both value-added measures for fifth-grade teachers continue to predict significant and persistent improvements in student achievement (columns III-IV and VII-VIII). For test scores, the effect falls from 0.19σ to 0.15σ when I restrict the sample to seventh to eighth-grade outcomes, which is consistent with teacher effects on test scores fading out over time, as documented by Chetty, Friedman, and Rockoff (2014a), Rothstein (2010), and Jacob, Lefgren, and Sims (2010). On the other hand, I do not observe the impacts of course grade value-added fading out comparably (0.18σ vs 0.19σ unconditionally).

I also find some limited evidence of multidimensional impacts, similar to Jackson (2018) and Petek and Pope (2023). Conditional on course grade value-added, teachers who are more effective in raising test scores also improve their students' grades in middle school by about 0.09σ (column VI). This effect shrinks to 0.04σ and becomes insignificant when restricting to seventh to eighth-grade outcomes, which is again consistent with the fade-out of test score effects. By contrast, fifth-grade teachers' value-added in course grades does not predict their students' middle school test scores (columns II and IV).

Table D.1 reports results separately for math and reading. The broad patterns are similar across subjects; both test score and course grade value-added are significant and persistent predictors of their respective outcomes. For math (Panel A), I find evidence of multidimensionality in both directions. Course grade value-added predicts students' math scores (about 0.05σ, Panel A, column II), and test score value-added predicts students' math grades (about 0.08σ, Panel A, column VI). Both effects are smaller and less precise when I restrict the outcomes to seventh and eighth grades (columns IV and VIII). For reading (Panel B), by contrast, the multidimensional effects are one-sided; test score value-added predicts reading grades (about 0.09σ and significant for sixth-eighth grade, and about 0.06σ and insignificant for seventh-eighth grade), but course grade value-added does not predict reading scores.

5.2 Gender Differentiated Results

Next, I investigate the heterogeneous impacts of fifth-grade teachers on boys and girls, by estimating Equation (7) for middle school test scores and course grades. Table 5 reports the gender-differentiated impacts of fifth-grade teacher value-added on middle school outcomes, averaged across subjects.

Table 5: Gender-Differentiated Impacts of Fifth-Grade Teachers on Middle School Outcomes

	Test Scores (6th-8th)		Test Scores (7th-8th)		Course Grades (6th-8th)		Course Grades (7th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Female	-0.042	-0.042	-0.143**	-0.143**	0.370***	0.370***	0.388***	0.385***
	(0.060)	(0.060)	(0.066)	(0.066)	(0.083)	(0.083)	(0.095)	(0.095)
Test Score VA	0.165***	0.160***	0.107***	0.102***		0.065		-0.020
	(0.033)	(0.033)	(0.035)	(0.036)		(0.048)		(0.054)
Female × Test Score VA	0.049	0.059	0.079	0.089*		0.048		0.109
	(0.046)	(0.047)	(0.049)	(0.050)		(0.065)		(0.074)
Course Grade VA		0.029		0.032	0.264***	0.246***	0.255***	0.260***
		(0.044)		(0.049)	(0.065)	(0.066)	(0.073)	(0.075)
Female × Course Grade VA		-0.066		-0.068	-0.141	-0.154*	-0.145	-0.176*
		(0.061)		(0.068)	(0.087)	(0.089)	(0.097)	(0.100)
Constant	-0.151	-0.152	-0.067	-0.068	-0.122	-0.120	-0.042	-0.041
	(0.182)	(0.182)	(0.205)	(0.205)	(0.181)	(0.181)	(0.218)	(0.218)
Mean	0.091	0.091	0.089	0.089	0.076	0.076	0.071	0.071
N	334425	334425	222950	222950	334425	334425	222950	222950
r^2	0.769	0.769	0.759	0.759	0.465	0.465	0.452	0.452

Note: Standard errors clustered at the Fifth-grade teacher level. Test Score VA is defined on fifth-grade test scores. Course Grade VA is defined on sixth-grade course grades. All outcomes are standardized (z-scores). Models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The results are consistent with the framework of relative skills developed in Section 4.1. Girls benefit more than boys from teachers with higher test score value-added, while boys benefit more than girls from teachers with higher course grade value-added. A 1σ increase in test score value-added improves boys' middle school test scores by about 0.16σ (column I). For girls, the effect is larger (roughly 0.21σ), though the differential effect is estimated imprecisely for sixth-eighth grades. When I restrict the outcomes to seventh-eighth grades, the female differential grows to about 0.08-0.09σ, and is statistically significant when I condition on course grade value-added (column IV). This is consistent with the prediction that teachers' cognitive-skill impacts are stronger for girls. By contrast, course grade value-added predicts improvements in middle school course grades of $\sim 0.25\sigma$ for boys and $\sim 0.1\sigma$ for girls (column VI). The differential improvement for boys is $\sim 0.15\sigma$) and is statistically significant. I observe a similar pattern when restricting outcomes to seventh-eighth grades. This implies that teachers who improve non-cognitive skills disproportionately benefit boys, which is again consistent with the relative skills interpretation.

Next, I report results from subject-specific versions of Equation (7), to test whether the patterns hold separately for math and reading, in Table 6. For math scores (reported in Panel A), a 1σ increase in a fifth-grade teacher's test score value-added improves boys' middle school math scores by about 0.12-0.13σ. For girls, the effect is larger, at $\sim$0.19-0.2σ, and the difference is statistically significant when the outcomes are restricted to seventh-eighth grades. This aligns with the model's prediction that teachers who improve cognitive skills disproportionately benefit girls. For math grades, boys' grades through middle school increase by about 0.21-0.24σ in response to a 1σ increase in their fifth-grade teacher's course grade value-added, and by $\sim 0.19\sigma$ when restricted to seventh and eighth grade. The corresponding effect for girls

is smaller, about 0.09-0.1σ, with the female interaction negative (though imprecisely measured). This indicates that teachers who improve non-cognitive skills generate higher gains for boys than girls, which is again consistent with the model's prediction. The effects are also directionally consistent for the impacts of fifth-grade teachers' test score value-added on math grades (girls improve more), and course-grade value-added on test scores (boys improve more).

Table 6: Gender-Differentiated Impacts of Fifth-Grade Teachers on Middle School Outcomes

Panel A: Math Outcomes								
	Math Scores (6th-8th)		Math Scores (7th-8th)		Math Grades (6th-8th)		Math Grades (7th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Female	0.162**	0.161**	0.218***	0.217***	0.438***	0.439***	0.455***	0.454***
	(0.072)	(0.072)	(0.080)	(0.080)	(0.094)	(0.094)	(0.108)	(0.108)
Test Score VA	0.138***	0.123***	0.089***	0.072**		0.072*		-0.008
	(0.027)	(0.028)	(0.029)	(0.030)		(0.038)		(0.043)
Female × Test Score VA	0.059	0.073**	0.084**	0.103**		0.013		0.047
	(0.036)	(0.037)	(0.040)	(0.041)		(0.052)		(0.059)
Course Grade VA		0.102**		0.117**	0.240***	0.211***	0.190***	0.193***
		(0.044)		(0.050)	(0.062)	(0.065)	(0.069)	(0.071)
Female × Course Grade VA		-0.099		-0.131*	-0.095	-0.099	-0.093	-0.112
		(0.063)		(0.069)	(0.085)	(0.087)	(0.094)	(0.097)
Constant	-0.186	-0.187	-0.073	-0.075	-0.336	-0.332	-0.383	-0.383
	(0.190)	(0.190)	(0.205)	(0.205)	(0.204)	(0.204)	(0.242)	(0.242)
Mean	0.099	0.099	0.098	0.098	0.077	0.077	0.072	0.072
N	334425	334425	222950	222950	334425	334425	222950	222950
r^2	0.707	0.707	0.699	0.699	0.389	0.389	0.377	0.377
Panel B: Reading Outcomes								
	Reading Scores (6th-8th)		Reading Scores (7th-8th)		Reading Grades (6th-8th)		Reading Grades (7th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Female	-0.250***	-0.250***	-0.511***	-0.511***	0.303***	0.304***	0.319***	0.318***
	(0.072)	(0.072)	(0.081)	(0.081)	(0.093)	(0.093)	(0.108)	(0.108)
Test Score VA	0.167***	0.167***	0.107*	0.106*		0.089		0.036
	(0.051)	(0.051)	(0.056)	(0.057)		(0.067)		(0.078)
Female × Test Score VA	0.032	0.035	0.061	0.064		0.000		0.061
	(0.071)	(0.072)	(0.079)	(0.080)		(0.092)		(0.106)
Course Grade VA		-0.002		0.006	0.193***	0.181***	0.235***	0.230***
		(0.047)		(0.053)	(0.065)	(0.066)	(0.075)	(0.076)
Female × Course Grade VA		-0.023		-0.020	-0.151*	-0.151*	-0.194*	-0.202**
		(0.067)		(0.074)	(0.086)	(0.087)	(0.100)	(0.101)
Constant	-0.032	-0.032	0.018	0.017	0.143	0.144	0.336	0.337
	(0.235)	(0.235)	(0.266)	(0.266)	(0.207)	(0.207)	(0.249)	(0.249)
Mean	0.082	0.082	0.080	0.080	0.075	0.075	0.070	0.070
N	334425	334425	222950	222950	334425	334425	222950	222950
r^2	0.675	0.675	0.664	0.664	0.387	0.387	0.381	0.381

Note: Standard errors clustered at the fifth-grade teacher level. Math and Reading Test Score VA are defined on fifth-grade test scores. Math and Reading Course Grade VA are defined on sixth-grade outcomes. All outcomes are standardized (z-scores). Models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Panel B of Table 6 reports fifth-grade teachers' heterogeneous impacts on boys' and girls' reading outcomes in middle school. Boys' reading scores in middle school increase by about 0.17σ with a 1σ increase in test score value-added of fifth-grade teachers, and this effect shrinks to $\sim$0.11σ for later grades. Girls' reading scores in middle school increase by about 0.2σ with a 1σ increase in test score value-added for fifth-grade teachers, and it reduces to $\sim$0.17σ in later grades. The difference, while not statistically significant, is directionally consistent with girls experiencing higher gains in response to their teachers

improving cognitive skills. Boys' reading grades increase by $\sim$0.18-0.19σ in response to a 1σ increase in course grade value-added of fifth-grade teachers, while girls' reading grades improve by about 0.03-0.04σ, implying that boys' reading grades improve considerably more than girls in response to their teachers improving non-cognitive skills (by about 0.15σ- a statistically significant effect). The difference is even higher in later grades ($\sim$0.19-0.2σ).

Taken together, these results provide clear evidence that teachers have gender-differentiated impacts that align with the relative skills interpretation. Specifically, teachers have stronger impacts in the dimension in which their students have a relative *deficiency*. Teachers with high test score value-added disproportionately improve girls' outcomes, while teachers with high course grade value-added disproportionately improve boys' outcomes. The gender-differentiated effects of test score value-added are strongest and most precisely estimated for math scores, and those of course grade value-added are strongest and most precisely estimated for reading grades.

5.2.1 Gender-Specific Value-Added Measures

To explore teachers' gender-differentiated impacts further, I estimate Equation (8), which uses both boy-specific and girl-specific teacher value-added, and allows for each value-added measure to predict outcomes separately for boys and girls.[12] This specification yields four sets of coefficients for each outcome, corresponding to the interaction between student gender and boy/girl-specific VA. Specifically, I test whether boy-specific and girl-specific value-added have cross-impacts across genders (in addition to predicting outcomes the focal gender) and whether those cross-impacts align with the predictions outlined in Section 4.1.

I report the results of these estimations in Table D.3 and Table D.4. For test scores, boy-specific value-added predicts improvements in middle school test scores for both boys and girls, while girl-specific value-added predicts improvements only for girls. In other words, both boys and girls benefit from teachers whose effectiveness is measured for their respective genders, and girls record additional improvements in response to their teachers' boy-specific effectiveness. For course grades, the pattern reverses. Girl-specific value-added in course grades predicts improvements in middle school course grades for both boys and girls, while boy-specific value-added only predicts improvements for boys. Here, boys benefit from teachers whose effectiveness is measured on either gender, while girls benefit only from teachers whose effectiveness is measured on girls. These patterns hold in both math and reading, and reinforce my earlier

[12] Boy (girl)-specific VA is estimated using just the set of boys (girls) in the sample. The procedure is described in greater detail in Section 4.2.2.

findings: teachers' effectiveness in improving to cognitive skills has stronger impacts on girls, while their effectiveness in improving non-cognitive skills has stronger impacts on boys.

5.3 Test of the Relative Skills Interpretation

The two-factor model developed in Section 4.1 provides a theoretical framework for the gender-differentiated teacher effects documented in Section 5.2. Specifically, the framework predicts that teachers improve students in the dimension in which they have a relative deficiency – so teachers who improve cognitive skills should disproportionately benefit students with a relative deficiency in cognitive skills, while those who improve non-cognitive skills should disproportionately benefit students with a relative deficiency in non-cognitive skills.

To test these predictions directly, I construct a student-level measure of relative proficiency using baseline academic performance. Specifically, I use fourth-grade test scores (y) and course grades (z) to classify students according to their relative strengths. For each student, I calculate the test score percentile within their cohort and construct a "score-grade equivalent" (q_y), by mapping this percentile to the grade distribution. I then define *score relative proficiency* as the difference between the score-grade equivalent and the actual grade point (Score RP=$q_y - z$).[13] Students with positive values of Score RP perform relatively better on test scores than grades, indicating a relative proficiency in test scores. Conversely, students with negative values of Score RP perform relatively better on grades than test scores, indicating a relative proficiency in course grades. Students with Score RP equal to zero do not exhibit any relative proficiency across the two measures. While this measure is necessarily discrete due to the limited variation in course grades, it provides an empirically tractable (though noisy) proxy for students' relative performance across the two dimensions of achievement.

If fifth-grade teachers' heterogeneous impacts on boys' and girls' middle school outcomes are indeed explained by this relative skills framework – then we should see teachers with higher test score value-added measures have stronger impacts for students with lower values of Score RP, and teachers with higher course grade value-added measures have stronger impacts for students with lower values of Score RP. To test this, I estimate the following model:

[13]First, I compute within-cohort percentiles for both fourth-grade test scores and fourth-grade course grades, separately for math and reading. Since course grades take only four discrete values (0=D or below, 1=C, 2=B, 3=A), I identify the percentile ranges corresponding to each grade point in the data. For example, if grades of D or below ($z = 0$) span the 1st to 10th percentile, C grades ($z = 1$) span the 11th to 36th percentile, B grades ($z = 2$) span the 37th to 71st percentile, and A grades ($z = 3$) span the 72nd percentile and above, I then map test score percentiles into these same ranges to create "score-grade equivalents" (q_y). A student whose test score falls in the1st-10th percentile receives $q_y = 0$, one in the 11th-36th percentile receives $q_y = 1$, and so on. Then, I define the measure of score relative proficiency as Score RP=$(q_y - z)$. I construct these measures separately for math and reading.

$$y^h_{icst} = \alpha_0 + \alpha_1 \delta^{VA}_j + \alpha_2 \text{Score RP}_i + \alpha_3 (\delta^{VA}_j \times \text{Score RP}_i) + \mathbf{Y}^{g-1}_{icst} \alpha_4 + \mathbf{X}_i \alpha_5 + \gamma^{g-1}_{cst} + \phi^g_s + \zeta_h + u^h_{icst} \quad (10)$$

Here, δ^{VA}_j denotes teacher j's value-added measure corresponding to outcome y^h_{icst}. When the outcome is a test score, $\delta^{VA}_j = \delta^{score}_j$, and when the outcome is a course grade, $\delta^{VA}_j = \delta^{grade}_j$. All other controls and fixed effects are as defined in Equation (9). Under the relative skills framework's predictions, $\alpha_3 < 0$ for test score outcomes and $\alpha_3 > 0$ for grade outcomes. In other words, teachers with high test score value-added should have larger effects on students with lower Score RP (i.e., those with a relative deficiency in test scores), while teachers with high grade value-added should have larger effects on students with higher Score RP (i.e., those with a relative deficiency in course grades).

Table 7 reports the results for my tests of this mechanism for seventh-eighth grade math scores and reading grades – the two outcomes with the strongest gender-differentiated impacts. Columns I and IV first report the gender-differentiated results from estimating Equation (7) without the relative proficiency measures-initially reported in Table 6. Columns II and IV then report the results from estimating Equation (10). The results provide some support for the relative skills interpretation. For math test scores (Column II), the interaction term between test score value-added and Score RP is negative and significant at the 10% level, consistent with the prediction that teachers with high test score value-added have larger effects on students with a relative deficiency in test scores. For reading grades (Column V), the interaction term between course grade value-added and Score RP has the predicted positive sign, indicating that teachers with high grade value-added have larger effects on students with a relative deficiency in course grades. However, this coefficient is not statistically distinguishable from zero. The point estimate is larger in magnitude than the corresponding coefficient for math scores, but the standard error is also substantially larger. This is consistent with course grades being inherently noisier measures than test scores, as I show in Section 3.3.

The main effects of Score RP are close to zero and statistically insignificant for both math scores and reading grades. This suggests that, conditional on the fourth-grade test scores and grades captured in the vector $\mathbf{Y}^{g-1}_{icst}$, the measure of relative proficiency itself does not independently predict middle school outcomes. The interaction effects therefore capture true heterogeneity in how different types of teachers affect students with different skill profiles, rather than simply reflecting differential trajectories by initial levels of relative skills.

Finally, Columns III and VI show the results for the gender-differentiated model (i.e., Equation (7)) when

Table 7: Test of the Relative Skills Interpretation:
Seventh-Eighth Grade Outcomes

	Math Scores			Reading Grades		
	I	II	III	IV	V	VI
Female	0.218***	0.056***	0.217***	0.319***	0.304***	0.320***
	(0.080)	(0.003)	(0.080)	(0.108)	(0.004)	(0.108)
TVA	0.089***	0.129***	0.089***	0.235***	0.142***	0.236***
	(0.029)	(0.021)	(0.029)	(0.075)	(0.050)	(0.075)
Score RP		0.000	-0.000		0.003	0.002
		(0.005)	(0.005)		(0.007)	(0.007)
Female × TVA	0.084**		0.079**	-0.194*		-0.188*
	(0.040)		(0.040)	(0.100)		(0.100)
Score RP × TVA		-0.040*	-0.035		0.062	0.054
		(0.024)	(0.025)		(0.064)	(0.063)
Constant	-0.073	-0.073	-0.072	0.336	0.376	0.337
	(0.205)	(0.206)	(0.205)	(0.249)	(0.250)	(0.250)
Mean	0.098	0.098	0.098	0.070	0.070	0.070
N	222,950	222,950	222,950	222,950	222,950	222,950
R^2	0.699	0.699	0.699	0.381	0.380	0.381

Note: Standard errors clustered at the fifth-grade teacher level. TVA refers to a 5th grade teacher's math score value-added (defined on 5th grade math scores) for columns I-III, and a 5th grade teacher's reading grade value-added (defined on 6th grade reading grades) for columns IV-VI. All outcomes are standardized (z-scores). Models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

the interaction of the value-added measure and the relative proficiency measure are added to it. If Score RP were measured without noise, and the gender-differentiated impacts were truly driven by relative deficiency – then the framework would predict that the gender-differentiated effects of high value-added teachers should be absorbed by the interaction term between Score RP and the value-added measure. While the gender-differentiated effects do not shrink meaningfully (by about 0.005σ for math scores and by 0.006σ for reading grades)- this could be due to measurement error in the construction of Score RP. I report the results for all four outcomes (math scores, math grades, reading scores, and reading grades) in Table D.5, and find similar results.

Taken together, these results suggest that the gender-differentiated teacher effects documented in Section 5.2 may indeed by operating through the relative skills interpretation outlined in Section 4.1. Teachers are indeed more effective at improving outcomes for students in dimensions where those students have a relative deficiency. Since boys on average have a relative proficiency in cognitive skills and girls in non-cognitive skills, this explains why teachers with high test score value-added disproportionately benefit girls, while those with high grade value-added disproportionately benefit boys.

6 Conclusion

Gender gaps in educational outcomes vary substantially across subjects and outcome types, with test score gaps and course grade gaps showing distinct patterns as students progress through school. The gender gap in course grades – which typically favors girls – persists (and indeed, expands) even after conditioning on test scores, suggesting that grades reward both non-cognitive and cognitive skills. Expanding gender gaps in grades (and, more generally, in non-cognitive skills) have profound consequences for future educational attainment: boys face higher dropout rates, lower graduation rates, and are less likely to enroll in college compared to girls. While prior research has established that teachers have persistent and distinct impacts on both cognitive and non-cognitive skills, whether and how these multidimensional teacher effects systematically differ for boys and girls remains unexplored – especially given the gender gaps in non-cognitive skills. In this paper, I address this gap by asking whether these dual dimensions of teacher quality have systematically different impacts for boys and girls – and what explains these gender-differentiated impacts.

Using administrative data from North Carolina, I estimate teacher value-added measures separately for test scores and course grades, and examine their gender-differentiated effects on their students' middle school outcomes. I find that teachers with high value-added in test scores disproportionately benefit girls (particularly in math), while teachers with high value-added in course grades disproportionately benefit boys (particularly in reading). These patterns are robust across specifications, and are reinforced by asymmetric cross-impacts in gender-specific value-added measures: boy-specific test score value-added persistently predicts test scores for both boys and girls, while girl-specific test score value-added only predicts outcomes for girls. Conversely, girl-specific grade value-added predicts course grades for both genders, while boy-specific grade value-added only predicts grades for boys. This pattern emerges in both math and reading, and is consistent with the gender-differentiated impacts that benefit girls in test scores and boys in grades.

These results align with a framework of relative skill differences that I describe in Section 4, in which test scores are relatively more intensive in cognitive skills and course grades are relatively more intensive in non-cognitive skills. Under this framework, observed gender gaps imply that boys have a relative proficiency in cognitive skills while girls have a relative proficiency in non-cognitive skills, and teachers improve students most in dimensions where they have a relative deficiency. This interpretation differs from existing work in teacher-value added literature, that emphasizes role-model effects or teacher bias in explaining heterogeneous impacts of teachers by gender (Barrios-Fernández and Riudavets-Barcons,

2024; García-Echalar, Poblete, and Rau, 2024). Evidence from Section 5.3 provides some support for this mechanism, though the discrete nature of the relative proficiency measure means that its precision is limited- and thus, any estimates of heterogeneous teacher effects by levels of relative skills should be considered as lower-bounds.

This paper contributes to literature on both gender gaps and teacher quality by documenting that teachers have gender-differentiated impacts that systematically vary across different types of outcomes. I argue that the observed gender-differentiated impacts reflect how teachers' strengths interact with students' baseline skill mixes. By connecting multidimensional teacher effectiveness to multidimensional gender gaps, these findings offer a different lens for understanding both phenomena. Understanding these heterogeneous impacts has implications for how we measure and interpret teacher effectiveness, and for how we think about the evolution of gender gaps through elementary and middle school.

A methodological contribution of my analysis is also to show that among non-cognitive measures, only course grades satisfy validation tests and persist over time, while absences, suspensions, and grade retention either fail standard validation checks or lack sufficient variation- as I demonstrate in Section 3.3. This suggests that course grade value-added should be treated as a standalone measure rather than combined with other behavioral outcomes in composite indices, such as those used by Jackson (2018) and Petek and Pope (2023).

Several caveats remain. The choice of fifth-grade teachers as the focal group of teachers is partly driven by the constraints of the data, but also because elementary school is when the same teacher teaches multiple subjects, allowing me to estimate teacher value-added measure across both math and reading for the same teachers. This is also the grade immediately before students transition to middle school, when gender gaps in grades expand substantially. For elementary and middle school, NCERDC reports anticipated rather than finalized grades, which introduces classical measurement error but still yields value-added measures that pass validation tests and persist through middle school. The relative proficiency measure I construct is necessarily discrete because course grades take only four values, which limits precision in the tests reported in Section 5.3.

Future work could test whether these patterns hold in other contexts or with more granular outcome measures (especially for grades), investigate the mechanisms more directly through classroom observations or experimental variation in teacher assignment, and explore implications for longer-run outcomes such as high school completion, college attendance, and track choice. Another potentially important direction is understanding whether teachers can be trained to improve both cognitive and non-cognitive skills simultaneously, and how student assignment policies might account for these multidimensional effects.

My paper documents a novel pattern in how teachers affect boys and girls differently across outcome types, and offers a relative skills framework for understanding it. The extent to which this framework explains these gender-differentiated impacts relative to other explanations, whether these patterns generalize to other settings, and what they mean for policy remain open questions.

References

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters (2017). Leveraging Lotteries for School Value-Added: Testing and Estimation. In: *The Quarterly Journal of Economics* 132.2, pp. 871–919. URL: https://doi.org/10.1093/qje/qjx001.

Aucejo, Esteban and Jonathan James (2021). The Path to College Education: The Role of Math and Verbal Skills. In: *Journal of Political Economy* 129.10, pp. 2905–2946. URL: https://doi.org/10.1086/715417.

Aucejo, Esteban M., Jane Cooley Fruehwirth, Sean Kelly, and Zachary Mozenter (2022). Teachers and the Gender Gap in Reading Achievement. In: *Journal of Human Capital* 16.3, pp. 372–403. URL: https://doi.org/10.1086/719731.

Aucejo, Esteban M. and Jonathan James (2019). Catching up to girls: Understanding the gender imbalance in educational attainment within race. In: *Journal of Applied Econometrics* 34.4, pp. 502–525. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2699.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman (2019). Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes. In: *American Economic Journal: Applied Economics* 11.3, pp. 338–81. URL: https://www.aeaweb.org/articles?id=10.1257/app.20170571.

Bacher-Hicks, Andrew, Thomas J Kane, and Douglas O Staiger (2014). Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles. Working Paper 20657. National Bureau of Economic Research. URL: http://www.nber.org/papers/w20657.

Barrios-Fernández, Andrés and Marc Riudavets-Barcons (2024). Teacher value-added and gender gaps in educational outcomes. In: *Economics of Education Review* 100, p. 102541. URL: https://www.sciencedirect.com/science/article/pii/S0272775724000359.

Bertrand, Marianne and Jessica Pan (2013). The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior. In: *American Economic Journal: Applied Economics* 5.1, pp. 32–64. URL: https://www.aeaweb.org/articles?id=10.1257/app.5.1.32.

Bettinger, Eric P. and Bridget Terry Long (2005). Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students. In: *American Economic Review* 95.2, pp. 152–157. URL: https://doi.org/10.1257/000282805774670149.

Card, David, Eric Chyn, and Laura Giuliano (2024). Can Gifted Education Help Higher-Ability Boys from Disadvantaged Backgrounds? Working Paper 33282. National Bureau of Economic Research. URL: http://www.nber.org/papers/w33282.

Card, David and A. Abigail Payne (2021). High School Choices and the Gender Gap in STEM. In: *Economic Inquiry* 59.1, pp. 9–28. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12934.

Carlana, Michela (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. In: *The Quarterly Journal of Economics* 134.3, pp. 1163–1224. URL: https://doi.org/10.1093/qje/qjz008.

Carrell, Scott E., Marianne E. Page, and James E. West (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap. In: *The Quarterly Journal of Economics* 125.3, pp. 1101–1144. URL: https://doi.org/10.1162/qjec.2010.125.3.1101.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. In: *American Economic Review* 104.9, pp. 2593–2632. URL: https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2593.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. In: *American Economic Review* 104.9, pp. 2633–79. URL: https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2633.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. In: *Journal of Human Resources* 41.4, pp. 778–820. URL: https://jhr.uwpress.org/content/41/4/778.

Cornwell, Christopher, David B. Mustard, and Jessica Van Parys (2013). Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School. In: *Journal of Human Resources* 48.1, pp. 236–264. URL: https://jhr.uwpress.org/content/48/1/236.

Dee, Thomas S. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? In: *American Economic Review* 95.2, pp. 158–165. URL: https://doi.org/10.1257/000282805774670446.

Delaney, Judith M. and Paul J. Devereux (2019). Understanding Gender Differences in STEM: Evidence From College Applications. In: *Economics of Education Review* 72, pp. 219–238. URL: https://www.sciencedirect.com/science/article/pii/S0272775719301761.

Fortin, Nicole M., Philip Oreopoulos, and Shelley Phipps (2015). Leaving Boys Behind. In: *Journal of Human Resources* 50.3, pp. 549–579. URL: https://jhr.uwpress.org/content/50/3/549.

Fryer Roland G., Jr. and Steven D. Levitt (2010). An Empirical Analysis of the Gender Gap in Mathematics. In: *American Economic Journal: Applied Economics* 2.2, pp. 210–40. URL: https://www.aeaweb.org/articles?id=10.1257/app.2.2.210.

García-Echalar, Andrés, Sebastián Poblete, and Tomás Rau (2024). Teacher value-added and the test score gender gap. In: *Labour Economics* 89, p. 102588. URL: https://www.sciencedirect.com/science/article/pii/S0927537124000836.

Gilraine, Michael and Odhrain McCarthy (2024). The Effect of the Prior Teacher on Value-Added. In: *The Review of Economics and Statistics*, pp. 1–45. URL: https://doi.org/10.1162/rest%5C_a%5C_01409.

Goldhaber, Dan, James Cowan, and Joe Walch (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. In: *Economics of Education Review* 36, pp. 216–228. URL: https://www.sciencedirect.com/science/article/pii/S0272775713000915.

Hanushek, Eric A. and Steven G. Rivkin (2010). Generalizations about Using Value-Added Measures of Teacher Quality. In: *American Economic Review* 100.2, pp. 267–71. URL: https://www.aeaweb.org/articles?id=10.1257/aer.100.2.267.

Heckman, James, Rodrigo Pinto, and Peter Savelyev (2013). Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. In: *American Economic Review* 103.6, pp. 2052–86. URL: https://www.aeaweb.org/articles?id=10.1257/aer.103.6.2052.

Heckman, James J., Jora Stixrud, and Sergio Urzua (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. In: *Journal of Labor Economics* 24.3, pp. 411–482. URL: https://doi.org/10.1086/504455.

Horváth, Hedvig (2015). Classroom Assignment Policies and Implications for Teacher Value-Added Estimation. Working Paper. URL: https://www.dropbox.com/scl/fi/ocomot9kfz3zlojgfo4vv/sorting_wp_horvath.pdf.

Jackson, C. Kirabo (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. In: *Journal of Political Economy* 126.5, pp. 2072–2107. URL: https://doi.org/10.1086/699018.

Jacob, Brian A. (2002). Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education. In: *Economics of Education Review* 21.6, pp. 589–598. URL: https://www.sciencedirect.com/science/article/pii/S0272775701000516.

Jacob, Brian A., Lars Lefgren, and David P. Sims (2010). The Persistence of Teacher-Induced Learning. In: *Journal of Human Resources* 45.4, pp. 915–943. URL: https://jhr.uwpress.org/content/45/4/915.

Kane, Thomas J and Douglas O Staiger (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Working Paper 14607. National Bureau of Economic Research. URL: http://www.nber.org/papers/w14607.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff (2015). Value-added modeling: A review. In: *Economics of Education Review* 47, pp. 180–195. URL: https://www.sciencedirect.com/science/article/pii/S0272775715000072.

Lavy, Victor and Rigissa Megalokonomou (2024). The Short- and the Long-Run Impact of Gender-Biased Teachers. In: *American Economic Journal: Applied Economics* 16.2, pp. 176–218. URL: https://www.aeaweb.org/articles?id=10.1257/app.20210052.

Lavy, Victor and Edith Sand (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. In: *Journal of Public Economics* 167, pp. 263–279. URL: https://www.sciencedirect.com/science/article/pii/S0047272718301750.

Lindqvist, Erik and Roine Vestman (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. In: *American Economic Journal: Applied Economics* 3.1, pp. 101–28. URL: https://www.aeaweb.org/articles?id=10.1257/app.3.1.101.

Martínez, Joan J. (2025). The Long-Term Effects of Teachers' Gender Stereotypes on Labor Outcomes. URL: https://joanjmartinez.nyc3.digitaloceanspaces.com/research/Martinez_LongTermEffects_Gender.pdf.

Mozenter, Zachary D. (2019). Essays on the Effects of Teacher Grading Standards and Other Teaching Practices. English. PhD thesis, p. 146. URL: https://www.proquest.com/dissertations-theses/essays-on-effects-teacher-grading-standards-other/docview/2240074332/se-2.

National Center for Education Statistics (2023). Status Dropout Rates. URL: https://nces.ed.gov/programs/coe/indicator/coj/status-dropout-rates.

National Center for Education Statistics (2024). Public High School Averaged Freshman Graduation Rate (AFGR), Digest of Education Statistics 2024. URL: https://nces.ed.gov/programs/digest/d24/tables/dt24_219.40.asp.

Penner, Andrew M. and Marcel Paret (2008). Gender Differences in Mathematics Achievement: Exploring The Early Grades and The Extremes. In: *Social Science Research* 37.1, pp. 239–253. URL: https://www.sciencedirect.com/science/article/pii/S0049089X07000427.

Petek, Nathan and Nolan G. Pope (2023). The Multidimensional Impact of Teachers on Students. In: *Journal of Political Economy* 131.4, pp. 1057–1107. URL: https://doi.org/10.1086/722227.

Pope, Devin G. and Justin R. Sydnor (2010). Geographic Variation in the Gender Differences in Test Scores. In: *Journal of Economic Perspectives* 24.2, pp. 95–108. URL: https://www.aeaweb.org/articles?id=10.1257/jep.24.2.95.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005). Teachers, Schools, and Academic Achievement. In: *Econometrica* 73.2, pp. 417–458. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00584.x.

Rothstein, Jesse (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. In: *The Quarterly Journal of Economics* 125.1, pp. 175–214. URL: https://doi.org/10.1162/qjec.2010.125.1.175.

Rothstein, Jesse (2017). Measuring the Impacts of Teachers: Comment. In: *American Economic Review* 107.6, pp. 1656–84. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20141440.

Staiger, Douglas O., Thomas J. Kane, and Brian D. Johnson (2024). Why Does Value-Added Work? Implications of a Dynamic Model of Student Achievement.

Terrier, Camille (2020). Boys lag behind: How teachers' gender biases affect student achievement. In: *Economics of Education Review* 77, p. 101981. URL: https://www.sciencedirect.com/science/article/pii/S0272775718307714.

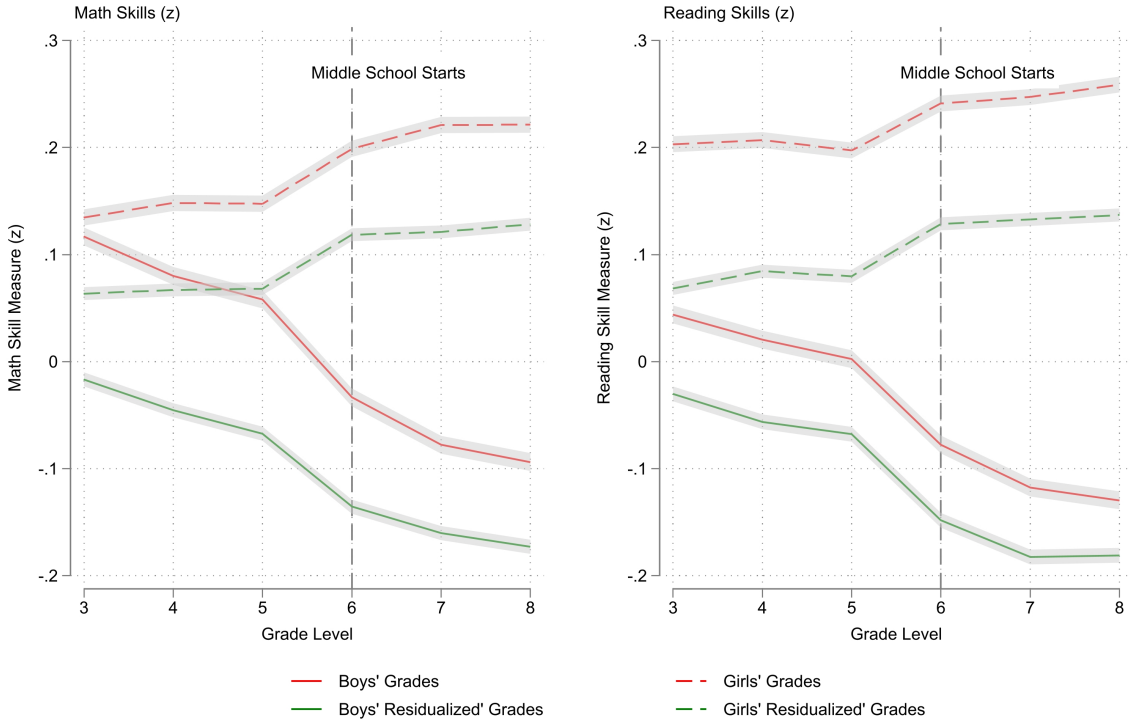
U.S. Bureau of Labor Statistics (2025). College Enrollment and Work Activity of Recent High School and College Graduates – 2024. URL: https://www.bls.gov/news.release/hsgec.htm (visited on 09/04/2025).

U.S. Department of Education, Office for Civil Rights (2023). Student Discipline and School Climate in U.S. Public Schools, 2020–21 CRDC. URL: https://www.ed.gov/sites/ed/files/about/offices/list/ocr/docs/crdc-discipline-school-climate-report.pdf (visited on 09/04/2025).

A Summary Statistics and Descriptive Relationships

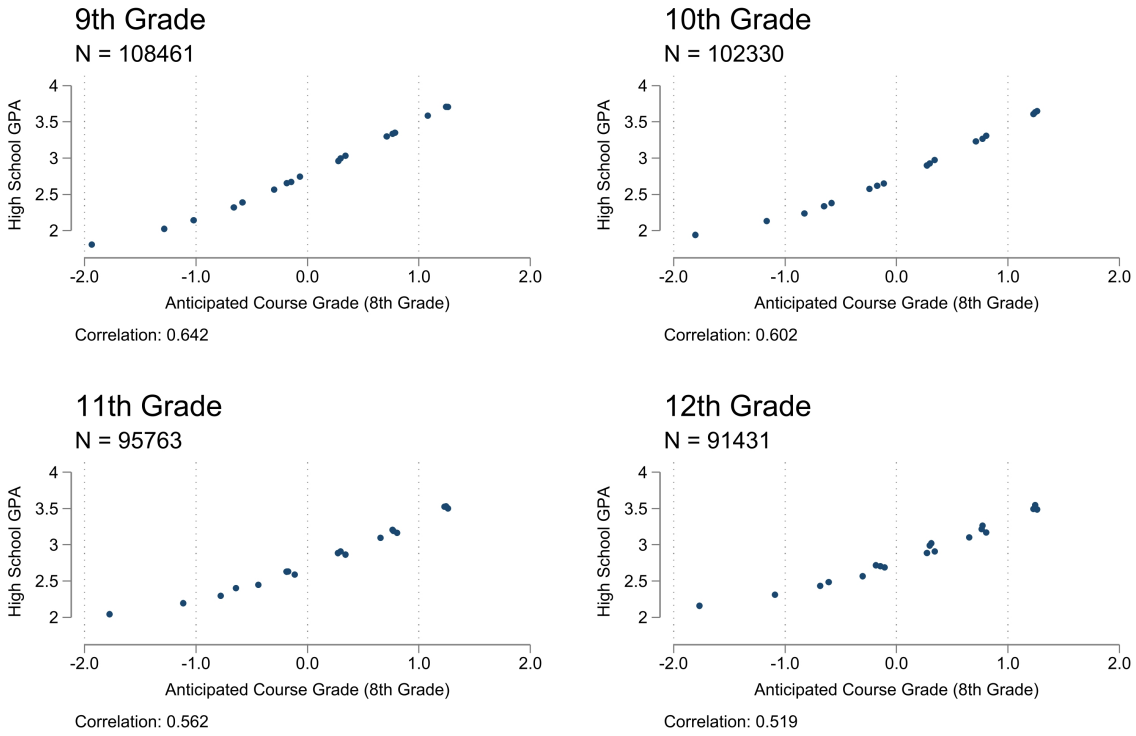
A.1 Progression of Test Scores and Course Grades

Figure A.1: Math and Reading Skills: Evolution of Gender Gaps
Third–Eighth Grade



Note. Test scores and course grades are standardized within grade-year. The sample is restricted to students who can be traced from third-eighth grade. Residualized grades are the residuals extracted from regressing course grades on test scores within each subject. I describe more details about the sample restrictions in Section 2.1.

Figure A.2: High School GPAs vs Eighth Grade Anticipated Course Grades
Ninth-Twelfth Grade



Note. Unweighted high school GPA's for each grade in high school plotted against the average (anticipated) grade for math and reading from eighth grade.

A.2 Sample Selection and Summary Statistics for Analysis Sample

Table A.1: Summary Statistics: Third-Grade Students (Unrestricted vs Analysis Sample

Student Outcomes			Student Characteristics		
	Mean	Diff		Mean	Diff
Test Score	-0.000	0.169***	Female	0.492	0.034***
	(1.000)	(0.004)		(0.500)	(0.002)
Math Score	0.001	0.181***	White	0.562	0.060***
	(1.000)	(0.004)		(0.496)	(0.002)
Reading Score	-0.000	0.169***	Black	0.254	-0.043***
	(1.000)	(0.004)		(0.435)	(0.002)
Course Grade	0.004	0.177***	Hispanic	0.114	-0.019***
	(0.947)	(0.003)		(0.318)	(0.001)
Math Grade	0.005	0.181***	Asian	0.020	-0.004***
	(0.997)	(0.004)		(0.139)	(0.001)
Reading Grade	0.003	0.173***	Other	0.050	0.006***
	(0.996)	(0.004)		(0.218)	(0.001)
Behavioral Skills	-0.002	0.196***	Disadvantaged	0.506	-0.048***
	(1.000)	(0.004)		(0.500)	(0.002)
ln(1+absences)	1.657	-0.078***	ESL	0.078	-0.024***
	(0.824)	(0.003)		(0.268)	(0.001)
Suspended	0.030	-0.017***	Reported Disability	0.123	0.065***
	(0.170)	(0.005)		(0.329)	(0.001)
Repeated Grade	0.007	-0.013***	Class Size	21.492	0.421***
	(0.084)	(0.003)		(4.701)	(0.017)

Note: Reported means are for the unrestricted sample. Standard deviations are reported in parentheses below the means. Differences are computed between students included in the analysis sample and those in the unrestricted sample who are not included. Standard errors for the difference in means are reported in parentheses below the differences. All test scores and grades are standardized (z-scores). Stars denote significance levels for a t-test for differences in means: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.2: Summary Statistics: Student Outcomes by Grade

	3rd Grade		4th Grade		5th Grade		6th Grade		7th Grade		8th Grade	
	Mean	Diff	Mean	Diff	Mean	Diff	Mean	Diff	Mean	Diff	Mean	Diff
Test Score (z)	0.114	-0.02**	0.110	-0.00	0.104	-0.02**	0.094	-0.02**	0.092	-0.06***	0.086	-0.07***
	(0.870)	(0.00)	(0.879)	(0.46)	(0.876)	(0.00)	(0.886)	(0.00)	(0.885)	(0.00)	(0.885)	(0.00)
Math Score (z)	0.119	0.10***	0.121	0.07***	0.119	0.07***	0.102	0.04***	0.102	-0.03***	0.094	-0.02***
	(0.943)	(0.00)	(0.947)	(0.00)	(0.949)	(0.00)	(0.955)	(0.00)	(0.957)	(0.00)	(0.963)	(0.00)
Reading Score (z)	0.109	-0.11***	0.100	-0.08***	0.090	-0.08***	0.086	-0.07***	0.083	-0.09***	0.078	-0.12***
	(0.939)	(0.00)	(0.944)	(0.00)	(0.946)	(0.00)	(0.947)	(0.00)	(0.950)	(0.00)	(0.952)	(0.00)
Course Grade (z)	0.126	-0.09***	0.116	-0.13***	0.103	-0.14***	0.086	-0.28***	0.073	-0.33***	0.069	-0.35***
	(0.892)	(0.00)	(0.894)	(0.00)	(0.893)	(0.00)	(0.877)	(0.00)	(0.880)	(0.00)	(0.879)	(0.00)
Math Grade (z)	0.126	-0.02**	0.115	-0.07***	0.104	-0.09***	0.086	-0.23***	0.076	-0.30***	0.068	-0.32***
	(0.942)	(0.00)	(0.954)	(0.00)	(0.963)	(0.00)	(0.967)	(0.00)	(0.972)	(0.00)	(0.979)	(0.00)
Reading Grade (z)	0.126	-0.16***	0.116	-0.19***	0.102	-0.19***	0.086	-0.32***	0.070	-0.36***	0.070	-0.39***
	(0.940)	(0.00)	(0.946)	(0.00)	(0.952)	(0.00)	(0.962)	(0.00)	(0.976)	(0.00)	(0.979)	(0.00)
Behavioral Skills Index (z)	0.137	-0.09***	0.127	-0.16***	0.113	-0.21***	0.099	-0.30***	0.088	-0.31***	0.081	-0.28***
	(0.878)	(0.00)	(0.889)	(0.00)	(0.910)	(0.00)	(0.907)	(0.00)	(0.923)	(0.00)	(0.948)	(0.00)
ln(1+Absences)	1.641	-0.01*	1.620	0.01	1.615	0.04***	1.680	0.06***	1.713	0.04***	1.750	-0.01
	(0.802)	(0.04)	(0.812)	(0.06)	(0.823)	(0.00)	(0.837)	(0.00)	(0.856)	(0.00)	(0.871)	(0.09)
Suspended	0.017	0.02***	0.034	0.04***	0.050	0.05***	0.090	0.08***	0.104	0.07***	0.106	0.06***
	(0.130)	(0.00)	(0.182)	(0.00)	(0.217)	(0.00)	(0.287)	(0.00)	(0.305)	(0.00)	(0.308)	(0.00)
Repeated Grade	0.012	0.00*	0.005	0.00**	0.004	0.00**	0.004	0.00***	0.004	0.00***	0.002	0.00***
	(0.108)	(0.03)	(0.069)	(0.00)	(0.062)	(0.00)	(0.061)	(0.00)	(0.060)	(0.00)	(0.060)	(0.00)
N	111,475		111,475		111,475		111,475		111,475		111,475	

Note: Reported means are for the analysis sample. Standard deviations are reported in parentheses below the means. Differences are computed between boys and girls in the analysis sample. Standard errors for the difference in means are reported in parentheses below the differences. All test scores and grades are standardized (z-scores). Stars denote significance levels for a t-test for differences in means: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.3 Summary Statistics: Teacher Characteristics

Table A.3: Summary Statistics: Teacher Characteristics

	Full Set	Analysis Sample
Female	0.88	0.87
	(0.33)	(0.34)
Black	0.14	0.14
	(0.35)	(0.34)
Hispanic	0.01	0.01
	(0.09)	(0.08)
White	0.85	0.85
	(0.36)	(0.35)
Experience (Years)	9.08	10.08
	(5.58)	(5.33)
Graduate Degree	0.34	0.35
	(0.48)	(0.48)
Observations	8,591	4,920

Note: Standard deviations in parentheses. The analysis sample is restricted based on the conditions described in Section 2.1.

B Summary Statistics and Descriptive Relationships for Value-Added Measures

B.1 Standard Deviations and Correlations

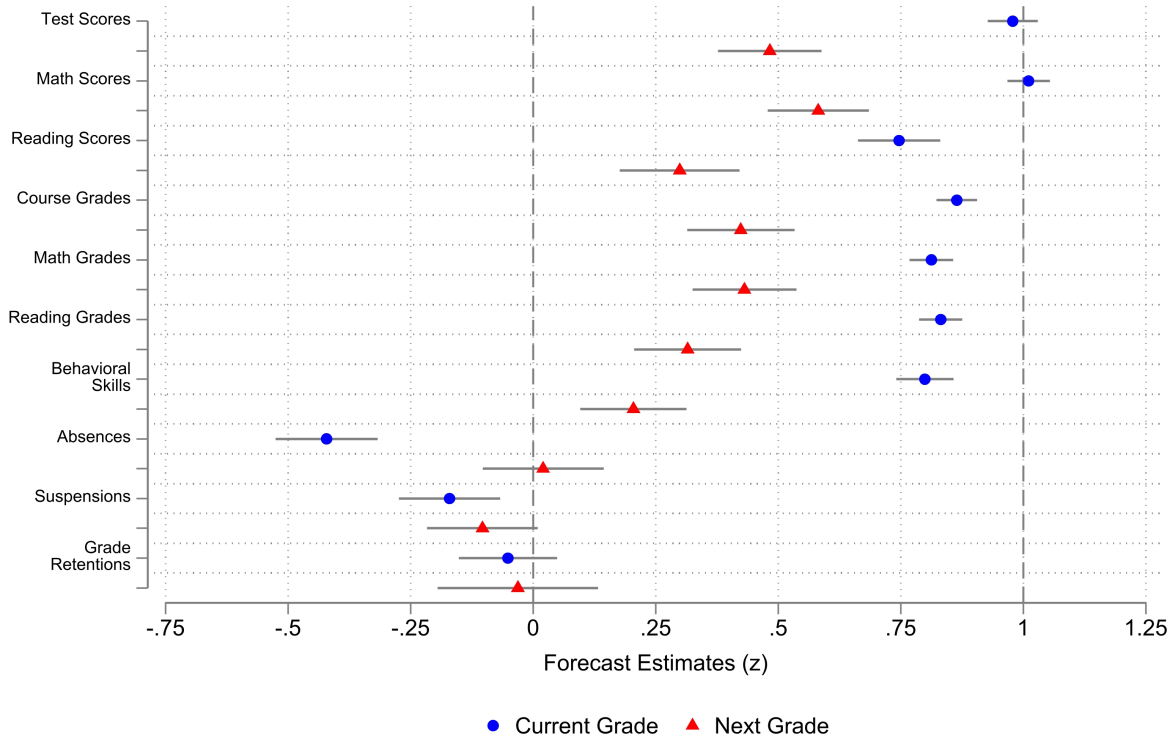
Table B.1: Correlations and Standard Deviations of Fourth-Grade Teacher Value-Added Measures

	Test Score	Math Score	Reading Score	Course Grade	Math Grade	Reading Grade	Behavioral Skills	Absences	Suspensions	Grade Repetition
Test Score	0.118									
Math Score	0.882	0.153								
Reading Score	0.801	0.470	0.124							
Course Grade	0.285	0.248	0.234	0.142						
Math Grade	0.286	0.288	0.188	0.850	0.159					
Reading Grade	0.230	0.165	0.243	0.848	0.533	0.161				
Behavioral Skills	0.189	0.166	0.150	0.609	0.536	0.547	0.171			
Absences	0.037	0.042	0.027	0.098	0.084	0.092	0.551	0.175		
Suspensions	0.002	-0.002	0.001	0.043	0.035	0.040	0.490	0.108	0.200	
Grade Repetition	0.003	-0.005	0.001	0.057	0.046	0.035	0.194	0.011	0.048	0.188

Note: The table reports correlations (below the diagonal) and standard deviations (on the diagonal) of teacher value-added measures.

B.2 Validation Tests for Fourth-Grade Teacher Value-Added Measures

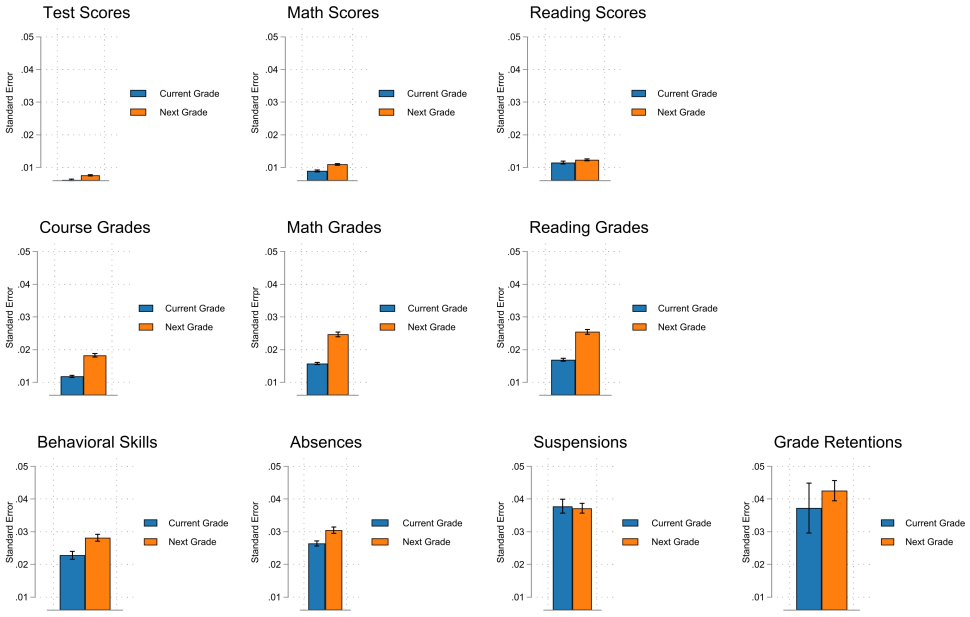
Figure B.1: Validations of Value-Added Measures
Fourth-Grade Teachers



Note. Figure reports estimated values of $\hat{\alpha_1}$ for value-added measures constructed for test scores, course grades, and behavioral outcomes, measured in standard deviation units of the outcomes. Estimated coefficients are from regressions wherein 4th (current) and 5th (next) grade outcomes are regressed on 4th grade VA measures.

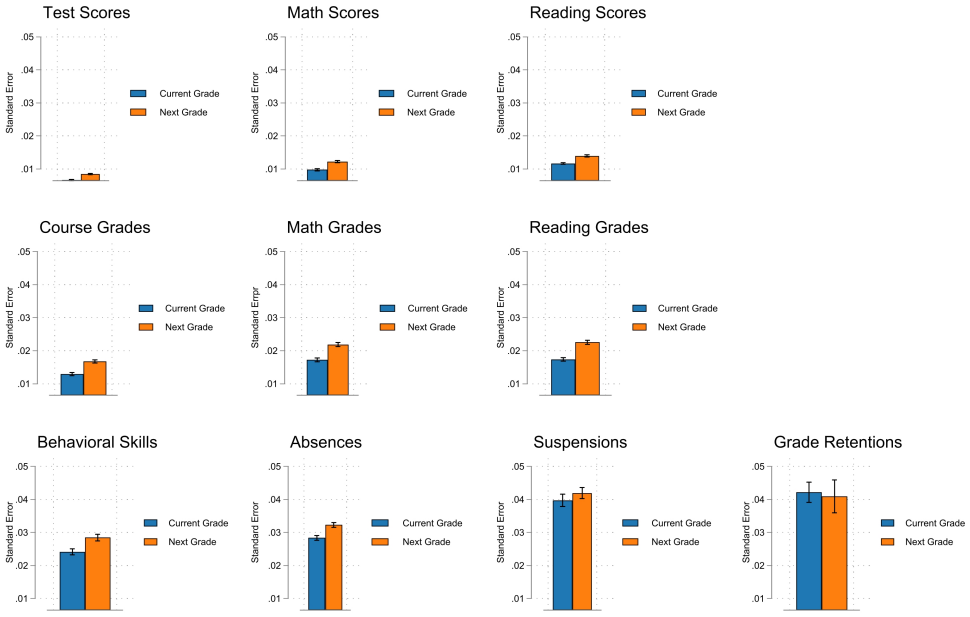
B.3 Standard Errors of Unshrunk Fixed Effects

Figure B.2: Standard Errors of Unshrunk Teacher Fixed Effects
Fifth-Grade Teachers



Note. Figure plots standard errors of unshrunk fixed effects for all value-added measures, constructed for contemporaneous ($h = g$, blue bars) and lead ($h = g + 1$, orange bars) outcomes, for 5th grade teachers. Higher means of standard errors indicate noisier estimates, and therefore greater shrinkage towards the mean.

Figure B.3: Standard Errors of Unshrunk Teacher Fixed Effects
Fourth-Grade Teachers



Note. Figure plots standard errors of unshrunk fixed effects for all value-added measures, constructed for contemporaneous ($h = g$, blue bars) and lead ($h = g + 1$, orange bars) outcomes, for 4th grade teachers. Higher means of standard errors indicate noisier estimates, and therefore greater shrinkage towards the mean.

B.4 Behavioral Skills Value-Added Regressions

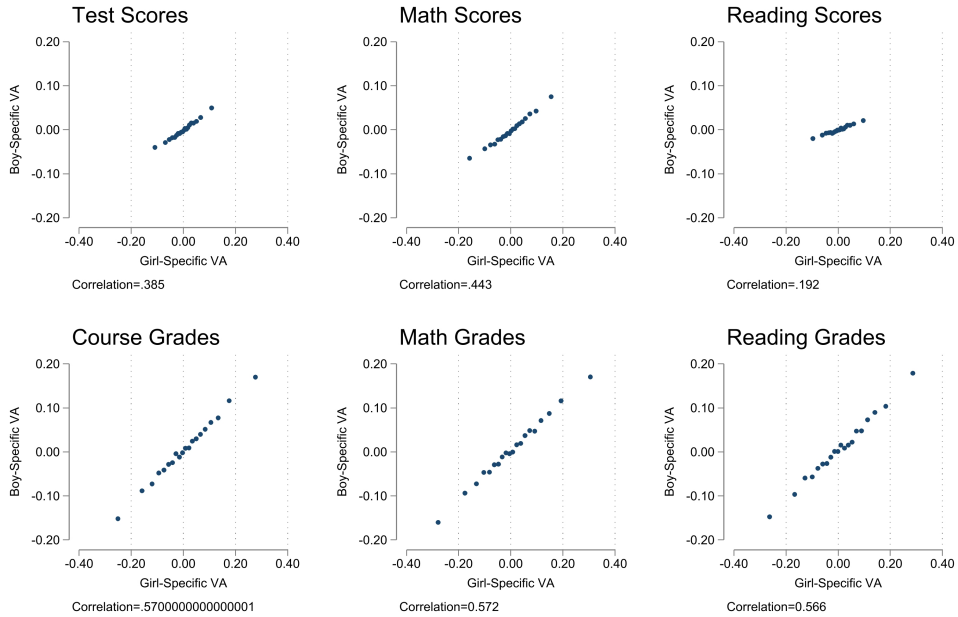
Table B.2: Multidimensional Impacts of Fifth-Grade Teachers on Middle School Behavioral Outcomes

	Absences (6th-7th Grade)			Suspended (6th-8th Grade)			Repeated Grade (6th-8th Grade)		
	I	II	III	IV	V	VI	VII	VIII	IX
Test Score VA			-0.112***			0.077**			0.012
			(0.042)			(0.038)			(0.025)
Course Grade VA		-0.037	-0.006		-0.047	-0.069		0.057*	0.053
		(0.057)	(0.058)		(0.051)	(0.052)		(0.034)	(0.035)
Behavioral VA	0.008	0.013	0.018	-0.036	-0.032	-0.029	-0.039	-0.044	-0.043
	(0.059)	(0.059)	(0.059)	(0.040)	(0.041)	(0.041)	(0.031)	(0.031)	(0.031)
Constant	-0.382	-0.381	-0.384	0.114	0.114	0.115	0.340*	0.339*	0.339*
	(0.268)	(0.268)	(0.268)	(0.357)	(0.357)	(0.357)	(0.195)	(0.195)	(0.195)
Mean	-0.048	-0.048	-0.048	-0.068	-0.068	-0.068	-0.039	-0.039	-0.039
N	222950	222950	222950	334425	334425	334425	334425	334425	334425
R^2	0.321	0.321	0.321	0.171	0.171	0.171	0.046	0.046	0.046

Note: Standard errors clustered at the fifth-grade teacher level. Test Score VA defined on fifth-grade test scores. Course grade and Behavioral VA measures defined on sixth-grade outcomes. Behavioral VA is Absences VA for columns I-III, Suspensions VA for columns IV-VI, and Grade Repetition VA for columns VII-IX. All outcomes are z-scores. All models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

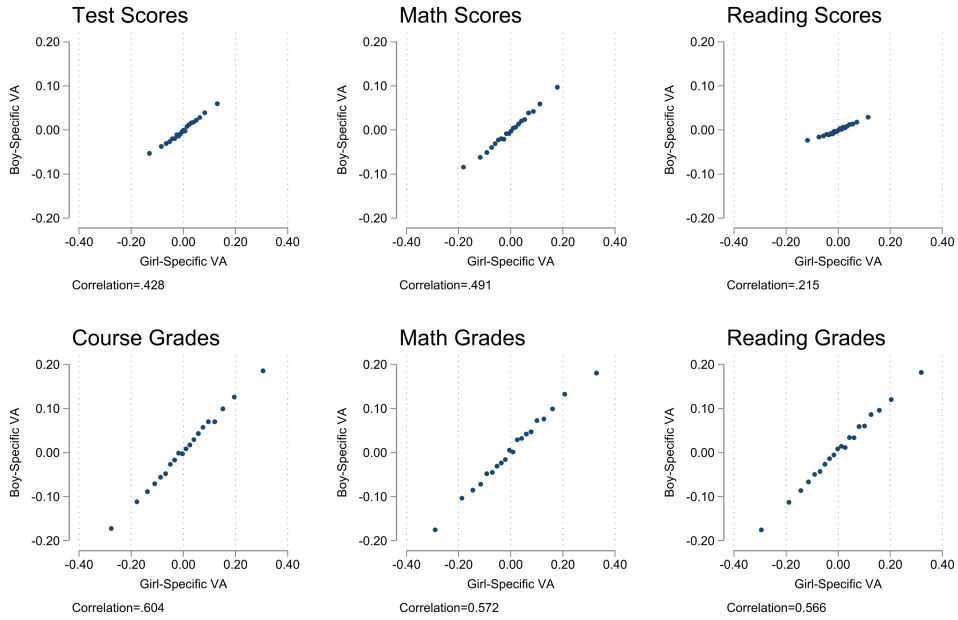
B.5 Gender-Specific Value-Added Measures: Descriptive Relationships

Figure B.4: Boy-Specific vs Girl-Specific Value-Added
Fifth-Grade Teachers



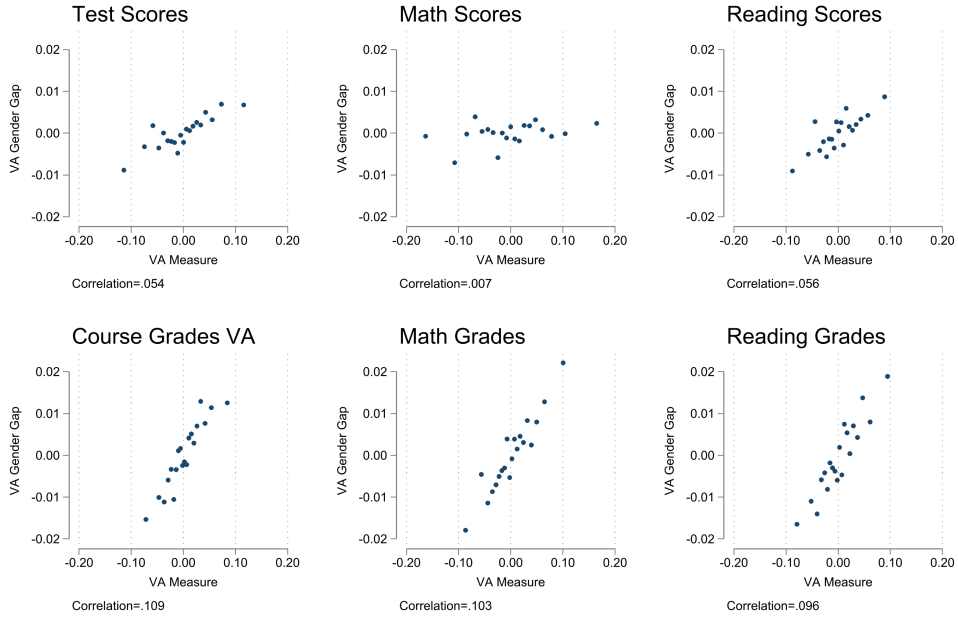
Note. Figure plots binned scatterplots of boy-specific value-added measures against girl-specific value-added measures of 5th grade teachers for test scores and course grades. Boy (girl)-specific value-added is defined as a teacher value-added measure based on only the boys (girls) taught by a 5th grade teacher- as described in Section 3.4

Figure B.5: Boy-Specific vs Girl-Specific Value-Added
Fourth-Grade Teachers



Note. Figure plots binned scatterplots of boy-specific value-added measures against girl-specific value-added measures of 4th grade teachers for test scores and course grades. Boy (girl)-specific value-added is defined as a teacher value-added measure based on only the boys (girls) taught by a 4th grade teacher- as described in Section 3.4

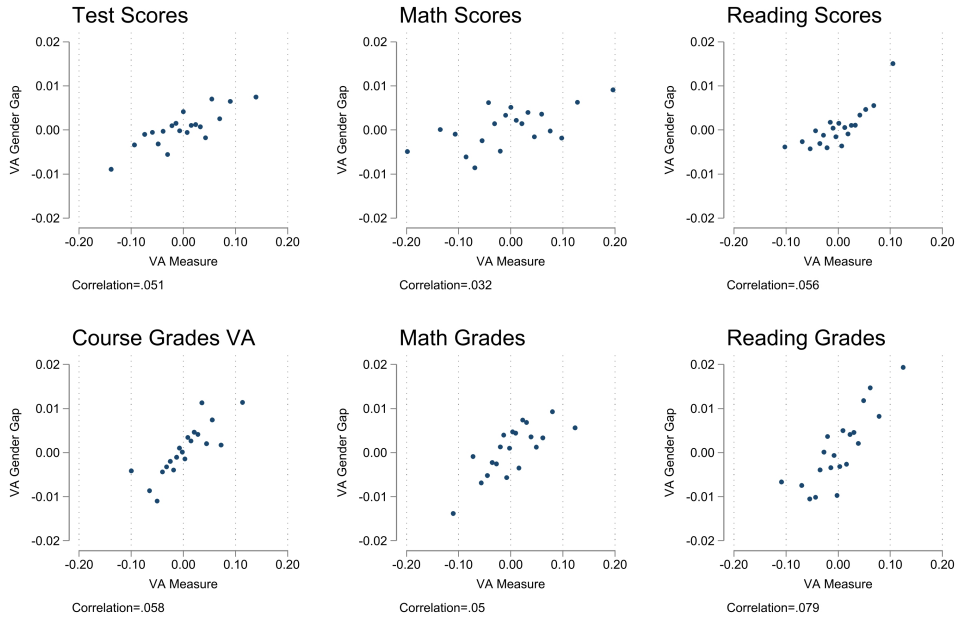
Figure B.6: Value-Added Gender Gaps
Fifth-Grade Teachers



Note. Figure plots binned scatterplots of value-added gender gaps against the overall value-added measures of 5th grade teachers for test scores and course grades. Value-added gender gaps are defined as the difference between a 5th grade teacher's boy-specific value-added measure and girl-specific value-added measure

Figure B.7: Value-Added Gender Gaps

Fourth-Grade Teachers



Note. Figure plots binned scatterplots of value-added gender gaps against the overall value-added measures of 4th grade teachers for test scores and course grades. Value-added gender gaps are defined as the difference between a 4th grade teacher's boy-specific value-added measure and girl-specific value-added measure

C Gender-Differentiated Impacts

Teachers can produce improvements in students' outcomes through their independent effects on cognitive skills (c) or non-cognitive skills (n). These effects are expressed by the following partial derivatives of test scores (y) and grades (z):

$$\frac{\partial y_{js}}{\partial c_{js}} = \theta \left(\frac{1}{\psi_{js}} \right)^{1-\theta}, \tag{C.1}$$

$$\frac{\partial z_{js}}{\partial c_{js}} = \gamma \left(\frac{1}{\psi_{js}} \right)^{1-\gamma}, \tag{C.2}$$

$$\frac{\partial y_{js}}{\partial n_{js}} = (1-\theta)\, \psi_{js}^{\theta}, \tag{C.3}$$

$$\frac{\partial z_{js}}{\partial n_{js}} = (1-\gamma)\, \psi_{js}^{\gamma}, \tag{C.4}$$

where $\psi_{js} = c_{js}/n_{js}$ denotes the relative skill mix for gender $j \in \{b, g\}$ in subject s.

Teachers Who Improve Cognitive Skills

1. On test scores:

$$\frac{\text{girls' improvement}}{\text{boys' improvement}} = \frac{\theta\left(\frac{1}{\psi_{gs}}\right)^{1-\theta}}{\theta\left(\frac{1}{\psi_{bs}}\right)^{1-\theta}} = \left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{1-\theta} > 1. \tag{C.5}$$

2. On grades:

$$\frac{\text{girls' improvement}}{\text{boys' improvement}} = \frac{\gamma\left(\frac{1}{\psi_{gs}}\right)^{1-\gamma}}{\gamma\left(\frac{1}{\psi_{bs}}\right)^{1-\gamma}} = \left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{1-\gamma} > 1. \tag{C.6}$$

$\Rightarrow$ Girls improve more.

Teachers Who Improve Non-Cognitive Skills

1. On test scores:

$$\frac{\text{boys' improvement}}{\text{girls' improvement}} = \frac{(1-\theta)\,\psi_{bs}^{\theta}}{(1-\theta)\,\psi_{gs}^{\theta}} = \left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{\theta} > 1. \tag{C.7}$$

2. On grades:

$$\frac{\text{boys' improvement}}{\text{girls' improvement}} = \frac{(1-\gamma)\,\psi_{bs}^{\gamma}}{(1-\gamma)\,\psi_{gs}^{\gamma}} = \left(\frac{\psi_{bs}}{\psi_{gs}}\right)^{\gamma} > 1. \tag{C.8}$$

$\Rightarrow$ Boys improve more.

D Results

D.1 Persistence of Fifth-Grade Teacher Effects

Table D.1: Multidimensional Impacts of Fifth-Grade Teachers on Middle School Math and Reading Outcomes

Panel A: Math Outcomes

	Math Scores (6th-8th)		Math Scores (7th-8th)		Math Grades (6th-8th)		Math Grades (7th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Test Score VA	0.168***	0.160***	0.131***	0.124***		0.078***		0.016
	(0.019)	(0.019)	(0.020)	(0.021)		(0.025)		(0.028)
Course Grade VA		0.051*		0.049	0.193***	0.162***	0.144***	0.138***
		(0.031)		(0.034)	(0.042)	(0.044)	(0.046)	(0.048)
Constant	-0.189	-0.189	-0.075	-0.075	-0.326	-0.323	-0.379	-0.378
	(0.191)	(0.191)	(0.206)	(0.206)	(0.203)	(0.203)	(0.243)	(0.243)
Mean	0.099	0.099	0.098	0.098	0.077	0.077	0.072	0.072
r^2	0.707	0.707	0.699	0.699	0.388	0.388	0.376	0.376

Panel B: Reading Outcomes

	Reading Scores (6th-8th)		Reading Scores (7th-8th)		Reading Grades (6th-8th)		Reading Grades (7th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Test Score VA	0.182***	0.185***	0.137***	0.138***		0.085*		0.064
	(0.034)	(0.034)	(0.038)	(0.038)		(0.044)		(0.050)
Course Grade VA		-0.016		-0.007	0.119***	0.108**	0.138***	0.129**
		(0.033)		(0.037)	(0.044)	(0.045)	(0.050)	(0.051)
Constant	0.000	0.000	0.053	0.053	0.170	0.171	0.375	0.375
	(0.235)	(0.235)	(0.266)	(0.266)	(0.209)	(0.209)	(0.250)	(0.250)
Mean	0.082	0.082	0.080	0.080	0.075	0.075	0.070	0.070
N	334425	334425	222950	222950	334425	334425	222950	222950
r^2	0.675	0.675	0.663	0.663	0.386	0.386	0.380	0.380

Note: Standard errors clustered at the fifth-grade teacher level. Test Score VA is defined on fifth-grade test scores. Course Grade VA is defined on sixth-grade course grades. All outcomes are standardized (z-scores). Models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.2 Persistence of Fourth-Grade Teacher Effects

Table D.2: Multidimensional Impacts of Fourth-Grade Teachers on Middle School Outcomes

Panel A: All Outcomes

	Test Scores (5th-8th)		Test Scores (6th-8th)		Course Grades (5th-8th)		Course Grades (6th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Test Score VA	0.177***	0.173***	0.138***	0.135***		0.123***		0.084***
	(0.019)	(0.020)	(0.021)	(0.022)		(0.026)		(0.029)
Course Grade VA		0.018		0.013	0.190***	0.132***	0.112***	0.072*
		(0.028)		(0.030)	(0.036)	(0.039)	(0.039)	(0.041)
Constant	-0.611***	-0.611***	-0.586***	-0.586***	-0.461***	-0.462***	-0.404***	-0.405***
	(0.102)	(0.102)	(0.108)	(0.108)	(0.124)	(0.124)	(0.144)	(0.145)
Mean	0.094	0.094	0.091	0.091	0.083	0.083	0.076	0.076
r^2	0.751	0.751	0.747	0.747	0.463	0.463	0.451	0.451

Panel B: Math Outcomes

	Math Scores (5th-8th)		Math Scores (6th-8th)		Math Grades (5th-8th)		Math Grades (6th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Test Score VA	0.189***	0.188***	0.144***	0.142***		0.130***		0.086***
	(0.016)	(0.016)	(0.017)	(0.018)		(0.020)		(0.023)
Course Grade VA		0.005		0.010	0.187***	0.109***	0.106***	0.054
		(0.030)		(0.032)	(0.036)	(0.038)	(0.039)	(0.041)
Constant	-0.555***	-0.555***	-0.556***	-0.556***	-0.471***	-0.475***	-0.469***	-0.472***
	(0.108)	(0.108)	(0.115)	(0.115)	(0.120)	(0.121)	(0.145)	(0.145)
Mean	0.104	0.104	0.099	0.099	0.083	0.083	0.077	0.077
r^2	0.689	0.689	0.687	0.687	0.391	0.391	0.378	0.378

Panel C: Reading Outcomes

	Reading Scores (5th-8th)		Reading Scores (6th-8th)		Reading Grades (5th-8th)		Reading Grades (6th-8th)	
	I	II	III	IV	V	VI	VII	VIII
Test Score VA	0.145***	0.135***	0.124***	0.117***		0.062*		0.046
	(0.031)	(0.032)	(0.033)	(0.034)		(0.035)		(0.040)
Course Grade VA		0.034		0.025	0.170***	0.153***	0.121***	0.109***
		(0.031)		(0.033)	(0.035)	(0.037)	(0.039)	(0.040)
Constant	-0.577***	-0.578***	-0.526***	-0.526***	-0.356**	-0.355**	-0.279*	-0.279*
	(0.120)	(0.120)	(0.126)	(0.126)	(0.147)	(0.147)	(0.169)	(0.169)
Mean	0.084	0.084	0.082	0.082	0.082	0.082	0.075	0.075
N	445900	445900	334425	334425	445900	445900	334425	334425
r^2	0.662	0.662	0.658	0.658	0.392	0.392	0.383	0.383

Note: Standard errors clustered at the fourth-grade teacher level. Test Score VA is defined on fourth-grade test scores. Course Grade VA is defined on fifth-grade course grades. All outcomes are standardized (z-scores). Models include third-grade classroom fixed effects and fourth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.3 Gender-Specific VA Results

Table D.3: Gender-Differentiated Impacts of Fifth-Grade Teachers on Middle School Outcomes

	Test Scores		Course Grades	
	I	II	III	IV
Male × Boy-Specific Test Score VA	0.076***	0.076***		-0.038
	(0.018)	(0.022)		(0.033)
Female × Boy-Specific Test Score VA	0.071***	0.090***		-0.001
	(0.017)	(0.021)		(0.029)
Male × Girl-Specific Test Score VA	0.000	0.005		0.016
	(0.018)	(0.023)		(0.034)
Female × Girl-Specific Test Score VA	0.047***	0.056***		0.034
	(0.017)	(0.022)		(0.030)
Male × Boy-Specific Course Grade VA		0.001	0.073**	0.097**
		(0.024)	(0.030)	(0.038)
Female × Boy-Specific Course Grade VA		-0.038	0.005	0.000
		(0.024)	(0.025)	(0.032)
Male × Girl-Specific Course Grade VA		-0.009	0.062**	0.057
		(0.026)	(0.031)	(0.040)
Female × Girl-Specific Course Grade VA		-0.019	0.067**	0.040
		(0.026)	(0.028)	(0.036)
Constant	-0.153	-0.153	-0.125	-0.125
	(0.182)	(0.182)	(0.181)	(0.182)
Mean	0.092	0.092	0.077	0.077
N	332883	332883	332883	332883
r^2	0.769	0.769	0.465	0.465

Note: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D.4: Gender-Differentiated Impacts of Fifth-Grade Teachers on Middle School Outcomes

	Test Scores		Course Grades	
	I	II	III	IV
Panel A: Math Outcomes				
Male × Boy-Specific Math Score VA	0.070***	0.060***		-0.016
	(0.020)	(0.023)		(0.033)
Female × Boy-Specific Math Score VA	0.094***	0.109***		0.041
	(0.018)	(0.022)		(0.029)
Male × Girl-Specific Math Score VA	0.021	0.018		0.018
	(0.019)	(0.023)		(0.032)
Female × Girl-Specific Math Score VA	0.050***	0.061***		0.012
	(0.018)	(0.022)		(0.029)
Male × Boy-Specific Math Grade VA		0.023	0.065**	0.072*
		(0.026)	(0.032)	(0.040)
Female × Boy-Specific Math Grade VA		-0.032	0.010	-0.020
		(0.025)	(0.027)	(0.033)
Male × Girl-Specific Math Grade VA		0.008	0.081**	0.071*
		(0.028)	(0.032)	(0.039)
Female × Girl-Specific Math Grade VA		-0.026	0.072**	0.054
		(0.027)	(0.030)	(0.037)
Constant	-0.187	-0.188	-0.334	-0.333
	(0.191)	(0.191)	(0.204)	(0.204)
Mean	0.101	0.101	0.077	0.077
N	332,883	332,883	332,883	332,883
R^2	0.707	0.707	0.389	0.389
Panel B: Reading Outcomes				
Male x Boy-Specific Reading Score VA	0.069***	0.063**		0.005
	(0.022)	(0.025)		(0.034)
Female x Boy-Specific Reading Score VA	0.051**	0.066***		-0.009
	(0.021)	(0.024)		(0.029)
Male x Girl-Specific Reading Score VA	0.014	0.014		0.023
	(0.022)	(0.026)		(0.037)
Female x Girl-Specific Reading Score VA	0.073***	0.060**		0.049
	(0.021)	(0.025)		(0.031)
Male x Boy-Specific Reading Grade VA		0.015	0.093***	0.086**
		(0.027)	(0.032)	(0.037)
Female x Boy-Specific Reading Grade VA		-0.035	0.015	0.013
		(0.027)	(0.028)	(0.032)
Male x Girl-Specific Reading Grade VA		-0.001	0.059*	0.043
		(0.029)	(0.033)	(0.040)
Female x Girl-Specific Reading Grade VA		0.030	0.076**	0.046
		(0.029)	(0.030)	(0.037)
Constant	-0.037	-0.039	0.130	0.132
	(0.235)	(0.235)	(0.207)	(0.207)
Mean	0.083	0.083	0.076	0.076
N	332883	332883	332883	332883
R^2	0.675	0.675	0.387	0.387

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D.4 Relative Proficiency Results

Table D.5: Relative Proficiency: Seventh-Eighth Grade Outcomes

Panel A: Math Outcomes						
	Math Scores			Math Grades		
	I	II	III	IV	V	VI
Female	0.218***	0.056***	0.217***	0.455***	0.307***	0.455***
	(0.080)	(0.003)	(0.080)	(0.108)	(0.004)	(0.108)
TVA	0.089***	0.129***	0.089***	0.190***	0.145***	0.190***
	(0.029)	(0.021)	(0.029)	(0.069)	(0.046)	(0.069)
CA in Scores		0.000	-0.000		-0.004	-0.005
		(0.005)	(0.005)		(0.008)	(0.008)
Female × TVA	0.084**		0.079**	-0.093		-0.091
	(0.040)		(0.040)	(0.094)		(0.094)
CA in Scores × TVA		-0.040*	-0.035		0.019	0.016
		(0.024)	(0.025)		(0.059)	(0.059)
Constant	-0.073	-0.073	-0.072	-0.383	-0.379	-0.383
	(0.205)	(0.206)	(0.205)	(0.242)	(0.243)	(0.242)
Mean	0.098	0.098	0.098	0.072	0.072	0.072
N	222,950	222,950	222,950	222,950	222,950	222,950
R^2	0.699	0.699	0.699	0.377	0.376	0.377
Panel B: Reading Outcomes						
	Reading Scores			Reading Grades		
	I	II	III	IV	V	VI
Female	-0.511***	0.020***	-0.511***	0.319***	0.304***	0.320***
	(0.081)	(0.003)	(0.081)	(0.108)	(0.004)	(0.108)
TVA	0.107*	0.131***	0.105*	0.235***	0.142***	0.236***
	(0.056)	(0.038)	(0.056)	(0.075)	(0.050)	(0.075)
CA in Scores		-0.009	-0.009*		0.003	0.002
		(0.005)	(0.005)		(0.007)	(0.007)
Female × TVA	0.061		0.052	-0.194*		-0.188*
	(0.079)		(0.079)	(0.100)		(0.100)
CA in Scores × TVA		-0.093*	-0.094*		0.062	0.054
		(0.049)	(0.049)		(0.064)	(0.063)
Constant	0.018	0.055	0.019	0.336	0.376	0.337
	(0.266)	(0.266)	(0.266)	(0.249)	(0.250)	(0.250)
Mean	0.080	0.080	0.080	0.070	0.070	0.070
N	222,950	222,950	222,950	222,950	222,950	222,950
R^2	0.664	0.663	0.664	0.381	0.380	0.381

Note: Standard errors clustered at the fifth-grade teacher level. For Panel A, TVA refers to a 5th grade teacher's math score value-added (defined on 5th grade math scores) for columns I-III, and a 5th grade teacher's math grade value-added (defined on 6th grade math grades) for columns IV-VI. For Panel B, TVA refers to a 5th grade teacher's reading score value-added (defined on 5th grade reading scores) for columns I-III, and a 5th grade teacher's reading grade value-added (defined on 6th grade reading grades) for columns IV-VI. All outcomes are standardized (z-scores). Models include fourth-grade classroom fixed effects and fifth-grade school fixed effects. Stars denote significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.