NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

AIR

Teacher Quality Gaps by Disability and Socioeconomic Status: Evidence from Los Angeles

Ijun Lai

W. Jesse Wood

Scott A. Imberman

Nathan Jones

Katharine O. Strunk

WORKING PAPER No. 228-0220 • February 2020

Teacher Quality Gaps by Disability and Socioeconomic Status: Evidence from Los Angeles

Ijun Lai
Michigan State University

W. Jesse Wood
Michigan State University

Scott A. Imberman
Michigan State University and NBER

Nathan Jones
Boston University

Katharine O. Strunk
Michigan State University

**** PRELIMINARY DRAFT****

**** PLEASE DO NOT CITE WITHOUT AUTHORS' PERMISSION ****

Contents

Acknowledgments

We are grateful to the Los Angeles Unified School District for providing the data necessary to conduct this research, and in particular to Vivian Ekchian, Sergio Franco, Bryan Johnson, Cynthia Lim, Patricia Pernin and Kathy Hayes for their partnership in developing the research agenda and Inocencia Cordova, Marilyn Fuller, Crystal Jewett, Joshua Klarin, Jonathan Lesser, Jacob Guthrie and Kevon Tucker-Seeley for their assistance in obtaining and understanding the data. This research was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of our funders or the institutions to which the authors are affiliated.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

Corresponding Author: Ijun Lai, laiijun@msu.edu.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

Teacher Quality Gaps by Disability and Socioeconomic Status: Evidence from Los Angeles
Ijun Lai, W. Jesse Wood, Scott A. Imberman, Nathan Jones, Katharine O. Strunk
CALDER Working Paper No. 228-0220
February 2020

Abstract

While the majority of students with disabilities (SWDs) receive instruction from general education teachers, little empirical work has investigated the ways in which these students have equitable access to high-quality teachers. We explore the differences in teacher quality experienced by SWDs and general education (GEN) students and how that access varies with school-level disadvantage by estimating SWD teacher quality gaps in the Los Angeles Unified School District. We examine several different indicators of teacher effectiveness (hiring scores, teacher experience, teachers' ratings on their observation-based performance evaluations, and value-added measures) for general education teachers who instruct both SWDs and general education (GEN) students. We find that SWDs are significantly more likely to have lower math VAM teachers than their GEN peers, and these gaps do not vary by school-level disadvantage. We find no differences on the other indicators of teacher effectiveness.

Introduction

The passage of the Individuals with Disabilities Act (IDEA) in 1975 was a watershed moment for the education of students with disabilities (SWDs). The act established that SWDs are to be provided with a "free and appropriate public education in the least restrictive environment." Today, roughly 6.4 million public school students in the U.S. receive special education services annually. Further, schools have made great strides in including SWDs in general education classrooms. As of 2015, 62.5% of all students with disabilities were being educated in a general education classroom for 80% or more of their school day (U.S. Department of Education, 2019). These percentages are even higher for high-incidence disability categories, with 69.5% of students with specific learning disabilities and 86.9% of students with speech and language impairments receiving instruction in the general education classroom most of the day.

The U.S. Supreme Court established an even higher standard for special education with its 2017 decision in *Endrew F. v. Douglas County School District RE-1*. Rather than requiring equal access to a free and appropriate education, the court stressed the need to ensure equitable outcomes for SWDs, who continue to lag behind their nondisabled peers in math and reading achievement (e.g., Chudowsky, Chudowsky, & Keber, 2009; Schulte & Stevens, 2015; Schulte, Elliott, Tindal, & Nese, 2016). Given the well-established importance of high-quality teachers for students' achievement and learning outcomes (e.g., Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), this raises critical questions about teacher quality gaps (TQGs) between students with disabilities in general education classrooms and their peers.

The teacher quality literature has repeatedly documented the unequal distribution of teachers—both across and within schools. For example, Goldhaber, Lavery, and Theobald

(2015) and Goldhaber, Quince, and Theobald (2018) establish that substantial gaps exist in teacher quality across the socioeconomic distribution; low-income students consistently have less access to highly qualified teachers. This is perhaps unsurprising since disadvantaged schools have more difficulty attracting and retaining high quality teachers (e.g., Boyd, Lankford, Loeb, & Wyckoff, 2005). Furthermore, research shows that even within schools, students are often sorted to teachers of varying quality based on their academic and behavioral histories (e.g. Kalogrides & Loeb, 2013; Lankford, Loeb, & Wyckoff, 2002).It is unclear based on existing research whether we should expect any between- or within-school teacher quality gaps between SWD in general education classrooms and their non-SWD peers. While much of the teacher quality gaps for students of color and low-income students stems from sorting across districts and schools (e.g. Goldhaber et al, 2015), we do not expect similar sorting patterns for students with disabilities, who appear to be relatively evenly distributed across schools. We hypothesize that TQGs, if there are any, will be concentrated in within-school sorting. Examining the distribution of SWDs and GEN students across the state of North Carolina, Gilmour and Henry (2018a) found that SWDs were clustered in classrooms in non-random ways; they were more likely to have classmates with lower prior academic performance than their GEN peers. However, it remains to be seen whether SWDs have differential access to high-quality teachers within schools. On the one hand, it is possible that SWDs could be seen as more "difficult" students to teach and consequently, more likely to be assigned to lower quality teachers (e.g. Clotfelter, Ladd, & Vigdor, 2016). At the same time, because in many states and districts there is a higher level of accountability or attention paid to the placement of and opportunities given to SWDs (see, for example, Swaak, 2020), it could be the case that districts proactively assign SWDs to particularly effective teachers. Indeed, Gilmour and Henry (2018b) find little evidence

of TQGs for SWDs overall in North Carolina, though they find some gaps for select disability subgroups. However, it is unclear whether this pattern will also hold in other contexts.

In addition, one important aspect of TQGs not addressed in Gilmour and Henry (2018b) is the question of how TQGs vary by school-level disadvantage. We hypothesize that teacher quality gaps in high poverty schools may be felt even more acutely by students with disabilities. Gilmour and Wehby (2019) demonstrate that the likelihood of teacher turnover increases with the number of SWDs in the classroom. Given the higher concentration of SWD in higher poverty schools, this finding suggests that SWDs in disadvantaged schools may be even less likely to access high-quality teachers than both non-SWDs in high poverty schools and SWDs in low poverty schools. However, to date, no studies have directly examined whether, within and across schools of varying income levels, TQGs exist across students with and without disabilities.

We join these two strands of research and ask the following research questions in the Los Angeles Unified School District (LAUSD) context:

1) *Does teacher quality vary across schools with differing degrees of disadvantage?*
2) Are there SWD vs non-SWD Teacher Quality Gaps?
3) Do SWD vs non-SWD gaps vary by school-level disadvantage?
4) Do Teacher Quality Gaps vary by specific disability type?

This paper makes two primary contributions to the literature on TQGs among SWDs. First, we include multiple quality indicators such as value-added measures (VAMs), teachers' ratings on their observation-based performance evaluations, hiring scores, and teacher experience (novice status). This allows us to examine whether or not TQGs exist in a different context (LAUSD), and if so, across an expanded range of quality measures. Second, we examine whether the small overall quality gaps found by Gilmour and Henry (2018) might mask school-level variation related to students' socioeconomic status. Previous research has shown that higher

poverty schools have greater difficulty attracting and retaining teachers (e.g. Hanushek, Kane, & Rivkin, 2004). If higher-quality teachers are sorting into schools with fewer low socioeconomic students, we might find that TQGs are exacerbated across schools within the same district. Therefore, it is important to understand how TQGs might differ across schools with different degrees of student disadvantage rather than just overall differences within a district or state.

Data

Context

The Los Angeles Unified School District (LAUSD) is the second largest district in the country with approximately 570,000 K-12 students and 24,000 K-12 teachers in 2018. The district is a particularly useful location in which to study TQGs as the district's administrative data provide detailed information on student and teacher characteristics, including student disability type, and allow for student-teacher matches. It is also an economically and racially diverse district that provides substantial variation in teacher quality, school level wealth and achievement, disability status, and student characteristics.

In general, LAUSD teachers are assigned to classes based on their credentials. At the elementary level, teachers may submit requests for track and grade level positions. Teachers may be assigned to their preferred classes based on District seniority, though principals may dispute specific assignments if they believe that the assignment is not in the best interest of the school. At the secondary level, teachers may submit requests for department selection, but principals assign teachers to specific classes and sections in consultation with department heads. See Appendix A for more details about teacher assignments.

Elementary classroom rosters are created at the end of the school year by grade-level teams. Since rosters are created at the end of the previous school year, this often means that they are created without any input from new teachers. After the start of school, grade-level teams can

call a meeting to ensure that students are equitably distributed across classrooms. If not, they can recommend changes to the principal. Student-teacher pairings for elementary school classes and core classes in middle school are "fairly randomized" and placement adjustments are mostly around balancing classroom sizes (LAUSD, personal communication, December 12, 2019).

In LAUSD, approximately 60% of students with mild or moderate disabilities spend the majority of their days in general education classrooms (Swaak, 2020). These placement decisions are made on a case-by-case basis by the child's Individualized Education Program team (composed of school personnel and relevant outside professionals), in collaboration with the child's parents. The decision is driven by the child's individualized needs and may result in a variety of placements, including in a general education classroom, in a self-contained or resource setting, or some combination of services. Among SWDs who are educated in the general education classroom, placement procedures do not differ from those for their non-disabled peers.

Sample

This study uses student- and teacher-level matched administrative data from SY2014-2015 through SY2017-2018. Data are provided by LAUSD's Office of Data and Accountability and the Division of Human Resources. Our sample includes all kindergarten through 8th grade students attending mainstream public schools during these years.[1] The data are at the student-year level and include demographic information such as disability status (detailed below), race/ethnicity, gender, free- or reduced-price lunch (FRL), and English Language Learner (ELL) status, as well as state standardized math and English Language Arts (ELA) test scores for students in grades three through eight. We normalize each subject's test scores to have a mean of zero and standard deviation of one for each grade-year combination. The data also contain teachers' demographic information (e.g. race/ethnicity, years of experience, and gender),

educational background (e.g. degrees obtained), and contract status (i.e. pre-tenure and permanent). Additionally, the teacher files include teachers' final evaluation scores as well as observation subcomponent scores and, for teachers hired since 2013-14, their hiring scores on a teacher screening system used by the district.

Students are linked to teachers through a transcript file, which provides details on students' classroom placements for each class period and the teacher of record. The final dataset is necessarily restricted to students who are linked to at least one teacher. While we have analyzed results for both math and ELA teachers, the results are very similar. Consequently, we focus our main results on math teachers and provide results for ELA teachers in Appendix Tables 1 and 3.

As we note above, our study focuses on SWDs who are taught by general education teachers since the majority of SWDs are in general education classrooms for most of their school days (U.S. Department of Education, 2019). While it may also be of interest to examine TQGs for SWDs taught in special education classrooms by special education teachers, data limitations make this problematic. In particular, we can only calculate VAM scores for approximately 15% of special education teachers (SETs) in our sample because students' previous test scores are used to construct VAMs and few SWDs with SETs have valid test scores from the previous year. Additionally, our VAMs are only constructed for teachers who have had at least eight students (with valid test scores) throughout the whole year. These constraints severely limit the number of SETs with VAMs. [2] For completeness, we calculate non-VAM TQGs for students in special education classrooms, broken down by school disadvantage level and disability type, in Appendix Tables 2 and 4. Our overall sample consists of 1,175,536 student-year observations, or 13,107 unique teachers in 619 schools.

Variables of interest

Disability Status

Using the detailed disability information in our data, we created indicator variables for four broad disability types—autism, specific learning disability, speech/language impairment, and other. The categories serve two purposes -- they reflect the disability groups that have high incidence rates in LAUSD and they reflect students with a range of needs. These four categories are not mutually exclusive since students may have multiple disabilities.[3]

School Characteristics

Since previous literature has shown that teacher quality can vary across schools with different characteristics, we generate school-level characteristics at the year-level and then average across the four years in our panel (SY2014-2015 through SY2017-2018). Our main analysis focuses on school-level free- or reduced-price lunch (FRL) status.[4] We split schools into three categories: less than 70% FRL, 70%-<95%, and 95%-100%. We chose these FRL categories based on a combination of how previous literature has examined the FRL distribution and the distribution of FRL students within LAUSD, which skews towards high rates of poverty.[5] Grouping schools in this manner allows us to compare students with and without disabilities at schools with similar demographics, while also observing how these differences may vary across schools with different student characteristics.

Teacher Characteristics

The literature suggests that teacher input variables, such as teachers' educational histories and credentials, are poorly correlated with teacher effectiveness (e.g., Angrist & Guryan, 2008; Chingos & Peterson, 2011; Clotfelter et al., 2007; Goldhaber & Brewer, 2000, 2001; Kane, Rockoff, & Staiger, 2008; Monk, 1994). As such, much of the more recent literature has

advocated for the use of teacher output measures as indicators of teacher quality, such as VAMs and teacher evaluation scores (e.g., Aaronson et al, 2007; Rivkin et al, 2005). Additional research has shown that exposure to early career teachers has negative impacts on student performance (Clotfelter, et al., 2007; Rice, 2010; Ladd & Sorensen, 2017; Staiger & Rockoff, 2010). Consequently, our main analysis focuses on four aspects of teacher quality: value-added measures (VAMs) of teachers' contributions to student achievement gains, teachers' ratings on their observation-based performance evaluations, teachers' initial hiring scores, and new teacher status (in first two years).

Teacher Value-Added Measures (VAM)

We calculate value-added measures (VAMs) for teachers teaching fourth through eighth grade. Following Chetty, Friedman, and Rockoff (2014), we use a multi-step calculation to create our value-added estimator. We begin by regressing test scores on student, classroom, and grade-level demographics to create residualized test scores. Specifically, we control for student-, classroom- and grade-level averages for race, gender, free/reduced lunch status, English Language Learner status, student with disability status, testing accommodation,[6] and previous test scores (cubed) in both the same subject and the "other" subject (i.e., to generate math VAMs, we include both lagged student test scores for math (same subject) and ELA (other subject)). Next, residualized test scores are averaged across all students for each teacher *j* in year *t.* We then calculate forecasting coefficients, which minimizes the mean squared error of the test-score predictions. Finally, data for teacher *j* in years outside of *t* are used to predict the value-added for teacher *j* in year *t*.[7] See Chetty et al., 2014 and Appendix B for a more detailed description.

Teacher Evaluation Scores

In LAUSD, all teachers new to a school are evaluated during their first two years through the Educator Development and Support: Teachers program, which focuses on classroom observations. After the first two years, teachers are evaluated at least every other year, but veteran teachers who meet certain qualifications may extend the time between evaluations to up to five years.[8] For each year that a teacher is evaluated, they are observed 1 or 2 times throughout that year and, depending on the year, received scores on between 7 and 15 subcomponents from the Teaching and Learning Framework (TLF), as well as an overall evaluation score. Three of these subcomponents are required for all teachers across all years, while other are selected by teachers before the observation period. We focus on the distribution of teacher scores across the TLF subcomponents.[9] Since observation components varied by academic year, and across teachers, we take the average score across all subcomponents and standardize by year.[10] Principals are encouraged to evaluate at least 25% of their teaching staff. In our sample, about 25-30% of our teachers are evaluated every year. Teachers who do not pass their evaluation are re-evaluated in the following year. We use teacher's evaluation scores from the prior year (or, for those who were not evaluated in the prior year, the most recent evaluation score before the current school year) to create our teacher evaluation measure.

Hiring Scores

LAUSD recently piloted (school year 2013-2014) and then fully adopted (2014-2015) a new teacher screening system. Consequently, we have hiring scores for teachers who were hired/re-hired since SY2014-2015. These are composite scores based on application information (such as licensure exam scores, grade point averages), professional references, writing sample, interview, and sample lesson demonstrations. More details can be found in Appendix C and Bruno and Strunk (2019). All hiring scores are standardized by year.

Experience

The current literature suggests there is a steep learning curve for novice teachers. On average, early career teachers rapidly improve their effectiveness over their first few years of teaching (e.g., Papay & Kraft, 2015; Kane et al., 2008; Rivkin et al., 2005), suggesting that new teachers are generally lower quality than more experienced teachers. Consequently, we examine students' exposure to novice teachers, which we define as having two or fewer years of experience.[11]

Overall Teacher Characteristics

Table 1 provides average general education teacher characteristics across FRL school bins, teacher quality measures. Panels A through D highlight that each teacher quality measure is coming from a different teacher subsample, with "Novice" teachers as the most inclusive sample. Specifically, only teachers who taught grades four through eight will have VAM scores, only teachers who have participated in the Educator Development and Support: Teachers program will have evaluation scores, and only teachers who have been hired (or re-hired) since SY2014-2015 will have valid hiring scores. To give a sense of how the sample varies across teacher quality measures, we also include descriptive statistics about the share of teachers with valid measures and the average score for these measures. For example, Panel A documents that about 32% of math teachers in our VAM sample have a valid teacher evaluation score and that the average z-scored evaluation score for this sample is 0.16.

There are many patterns that are consistent across all these subsamples. Most notably, students at lower FRL (i.e., higher income) schools tend to have more white and female teachers than students at higher FRL schools. Additionally, on average, students in lower FRL schools

tend to be exposed to teachers with higher teacher evaluation scores and hiring scores than students attending higher FRL schools.

Methods

We examine average teacher characteristics and whether these differ by student disability status. Our analytic approach is similar to that used in previous literature on TQGs (e.g., Clotfelter et al., 2005; Goldhaber et al., 2015). Specifically, we use a simple bivariate regression of the following form:

$$Y_{ijsb} = \beta_0 + \beta_1 Disability_{ijsb} + \varepsilon_{ijsb} \quad (1)$$

where Y_{ijsb} represents the teacher quality measure of interest (i.e. <= 2 years of experience) for student *i* matched with teacher *j* at school *s*, and *b* represents the disadvantage bin (i.e. FRL < 70%, FRL 70-<95%). For our main results, $Disability_{ijsb}$ is an indicator variable for students with disabilities. For our subgroup analysis, *Disability* represents one of three disability subgroups (Specific Learning Disability, Autism, or Speech/Language Impairment)[12] and zeros are given for non-SWDs. Standard errors are clustered at the school level. The bivariate regression allows us to calculate exposure rates to students with disabilities for teachers across different quality measures, as well as the exposure rates for their non-disabled peers, and to test if this difference (captured in β_1) is statistically significant. We also run a school fixed effects model, to test the stability of our coefficients. The addition of school fixed effects focuses the coefficients on within-school variation. Since results across the models are similar, we conclude that much of the TQGs are driven by within-school sorting, and only report coefficients from the unadjusted model (equation 1). However, we include information about between and within variance estimates from the school fixed effects model in Tables 3 and 4. The variance decompositions

are particularly interesting as they allow us to estimate the proportion of variation in TQGs that occurs within and across schools.

In addition to TQGs within FRL bins, we are interested in whether TQGs differ across bins (referred to as "disadvantage gaps" from this point forward). Specifically, we evaluate whether any of the TQGs are significantly different from the TQG in schools with less than 70% FRL students. To do this, we pool observations across two bins (with the most advantaged school as the reference bin) and estimate the following equation:

(2) $Y_{ijs} = \beta_0 + \beta_1 Disability_{ijs} + \beta_2 Adv\ Sch_{ijs} + \beta_3 (Disability_{ijs} * Adv\ Sch_{ijs}) + \varepsilon_{ijs}$

where, again, Y_{ijs} represents the teacher quality measure of interest (i.e. <= 2 years of experience) for student *i* in teacher *j* at school *s* and $Disability_{ijs}$ is an indicator variable for students with disabilities. $Adv\ Sch_{ijs}$ is an indicator variable for the most advantaged school (FRL <70%). The $Disability_{ijs}$*$Adv\ Sch_{ijs}$ interaction measures the teacher quality gap differences between the two school disadvantage bins and tests whether this difference is statistically significant. In the interest of space, we only display the p-value associated with $Disability_{ijs}$*$Adv\ Sch_{ijs}$ in our tables.

Results

Research Question 1: Does teacher quality in LAUSD vary across schools with differing degrees of disadvantage?

Previous literature has documented that schools with greater shares of students in poverty have, on average, lower-quality teachers (e.g., Clotfelter et al 2007; Goldhaber et al 2015; Goldhaber et al 2018; Sass et al 2012). Table 2 presents the mean and standard deviation for each general education math teacher characteristic within our three FRL bins. Consistent with

previous studies, we generally find increasing exposure to lower quality teachers as we move down the column from most to least advantaged schools. For example, the average teacher evaluation score for students attending the most advantaged schools in our sample (<70% FRL) was 0.361 (measured in standard deviation units), while the average score at the least advantaged school (>=95% FRL) was 0.161. We find significant disadvantage gaps (<70% FRL schools versus middle and highest FRL schools) for VAMs (lowest FRL vs middle FRL only) and teacher evaluation scores (across both FRL bins). However, we find no evidence of significant differences by FRL in terms of hiring score and novice teachers. This suggests that students in higher-poverty schools are taught by lower quality teachers in terms of their VAM (for the middle FRL group only) and teacher evaluation scores, but that, in contrast to studies in other contexts (e.g. Boyd et al, 2008; Clotfelter et al, 2007), novice teachers are relatively equitably distributed across LAUSD schools regardless of school-level disadvantage. In addition, we find that teachers in lower-poverty schools have lower initial hiring scores than do teachers in the other disadvantage bins, but these are not significant at traditional significance levels.

Research Question 2: Are there SWD vs non-SWD TQGs? Do these gaps vary by school-level disadvantage?

Table 3 presents overall SWD vs non-SWD TQGs. We begin by examining the average teacher quality for SWD in general education classrooms, and then the average teacher quality for their non-SWD peers, for each of our teacher quality measures- VAM, teacher evaluation scores, hiring scores, and novice teacher. Rows three and four present the quality gap and the corresponding standard error. We find that, relative to non-SWDs in general education math classrooms, SWDs with general education teachers (GET) are assigned to lower quality math teachers in terms of VAMs and teacher evaluation scores. On average, SWDs in general education classrooms have math teachers with 0.024 standard deviation lower VAMs and 0.028

lower standardized teacher evaluation scores than their non-SWD peers. There are no significant gaps in the experience (novice status) or hiring score of GETs teaching SWDs relative to non-SWDs.

The last two rows of Panel A present estimates from a model that adds a school fixed effect. This approach allows us to examine how much of the variance in the TQGs are due to within or between school factors. For the VAM, teacher evaluation, and novice measures, approximately 2/3 of the gaps are driven by within-school differences, suggesting that the gaps are mostly a function of within-school distribution of teachers to SWDs and non-SWDs, rather that teacher sorting across schools. For hiring scores there appears to be a larger role for factors that differ across schools, but 53% of the variance remains within-school. One possible explanation for this difference may be that higher turnover rates at certain schools are driving the increase in across-school variation for this teacher quality measure.

Tables 2 and 3 show that overall, there are TQGs by school-level disadvantage and, for SWD in general education classrooms, significant differences by disability status in average teacher VAMs and teacher evaluation scores. However, these findings are unable to shed light on how these factors interact. The rest of the paper explores how teacher quality varies when we examine student disability status and school poverty levels simultaneously.

Research Question 3: Do SWD vs non-SWD gaps vary by school-level disadvantage?

Table 4 presents the mean teacher quality scores and quality gaps across disability status and school disadvantage, with each column representing a different teacher characteristic of interest. Panel A presents the results for the most advantaged (<70% FRL) bin. We find significant SWD versus non-SWD TQGs across one measure: VAM (-0.047 standard deviations).

Panels B and C present our findings for the middle (70-<95% FRL) and most disadvantaged (>=95% FRL) schools. As in panel A, we see that SWD are more likely to have teachers with lower VAMs (Panel B: -0.014, Panel C: -0.019), though this difference is only significant at the highest poverty schools. We find no evidence of significant TQGs based on evaluation scores, hiring scores or novice status. The last row of Panels B and C displays the p-value testing for the disadvantage gap (comparing TQGs from each bin to the lowest FRL bin). Our estimates suggest that VAM (and all other) TQGs are similar across FRL bins.

Table 4 also provides between and within-school variance decomposition of the gaps within disadvantage bins. In general, the ratio of within- to between-school variance is similar across bins—though there are two notable exceptions. While hiring scores are relatively evenly split in terms of within- and between- school factors for the bottom and top FRL bins, the middle FRL bin is influenced more by between school factors (61%). While the hiring TQG is not significant for this FRL bin, our variance decomposition suggests that any differences that arise are more due to between-school sorting. The second exception comes from novice teachers. Within lower poverty schools, TQGs for novice teachers are mostly driven by within-school sorting (96%), while TQGs in the highest poverty schools are more evenly split (between-school variation: 40%, within-school variation: 61%).

Appendix Table 1 provides qualitatively similar estimates for ELA. For ELA the VAM gaps for the lowest poverty school is much smaller in magnitude and statistically insignificant while the teacher evaluation gap is larger and significant for the lower and middle FRL bins (-0.085 and -0.072 standard deviations respectively). Additionally, we find evidence of significantly greater exposure to novice teachers for students in the middle and highest poverty schools. These estimates are similar to the ones found in Table 4. Appendix Table 2 presents the

same results as in Table 4, but for SWD with SETs compared to non-SWD (with the exception of VAMs, which are not shown due to sample size constraints). The only teacher quality measure that is consistently significant across all FRL bins is teacher experience. SWDs with SETs are significantly more likely to have a novice math teacher than their non-SWD peers. This is not particularly surprising since many school districts are facing SET shortages, and consistently hiring more SETs. Additionally, we find that the TQG for the most disadvantaged schools is significantly greater than that at the most advantaged schools (p-value 0.02), suggesting greater inequality (in terms of teacher experience) for SWDs with SETs in high poverty schools.

Research Question 4: Do TQGs vary by specific disability type?

Looking across all students with disabilities may mask heterogeneous differences. Consequently, we disaggregate our data to more closely examine the three largest disability subgroups (ordered by prevalence): specific learning disability (SLD, ~55% of SWDs), autism (~18% of SWDs), and speech/language impairment (SLI, ~15% of SWDs). Table 5 presents our TQG estimates with these subsamples. Panel A presents the estimates for students with SLD, while panels B and C present the estimates for students with autism and SLI, respectively. Within each panel, we present the TQG (for each specific disability compared to non-SWD), standard errors, and sample size for each cell. Following the format in Table 4, we also include p-values for disadvantage gaps, which measure if TQGs in each FRL bin are significantly different from the TQG in the most advantaged schools (<70% FRL).

Results for students with SLD follow a similar pattern to the overall sample (shown in Table 4). Across the lowest and highest FRL bins, students with SLD have teachers with significantly lower VAMs (ranging from -0.029 to -0.074 standard deviations) compared to their peers without disabilities. Our estimates also suggest that students with SLD tend to have

teachers with lower evaluation scores, although these differences are only significant for the middle (-0.087 standard deviations) and highest FRL bins (-0.053 standard deviations). Like our main findings, we find no significant differences in terms of hiring scores and novice teachers. Interestingly, we find that the VAM TQGs in the lowest poverty bin is significantly greater than the VAM TQGs in the more disadvantaged school groups. TQGs across other teacher quality measures did not vary across FRL bins, suggesting little correlation between school-level disadvantage and SWD TQGs.

Estimates for students with autism and SLI suggest few significant differences between these subgroups and their non-SWD peers. If anything, our estimates suggest that depending on the FRL bin, these subgroups may be accessing higher quality teachers than their non-SWD peers. For example, students with SLI in the middle and highest FRL bins have teachers with significantly higher VAMs (0.076 and 0.058 standard deviations, respectively) than their non-SWD peers. Additionally, SLI in the highest FRL schools are more likely to access teachers with significantly higher evaluation scores (0.076 standard deviations) and less likely to have novice teachers. Overall, we generally do not find evidence that TQGs for students with autism or SLI vary by school disadvantage. The one exception is that the SLI teacher evaluation TQG between the most disadvantaged schools (a significant 0.076 difference) is significantly larger than the SLI TQG in the lowest FRL schools (a non-significant -0.021 standard deviation difference).

Appendix Table 3 displays the estimates for ELA teachers. Like the findings in Table 5, SLD students followed the same pattern for overall ELA teacher quality differences. We find no evidence of teacher quality gaps for students with autism and we find evidence that SLI students are more likely to be exposed to higher quality teachers than their non-SWD peers.

Appendix Table 4 presents the same estimates for SWD with SETs compared to their non-SWD peers, broken down by specific disability subcategories. Again, we find that much of the significant TQGs are concentrated within the Specific Learning Disability subgroup. Students with SLD in the middle and highest FRL bins have math teachers with significantly lower teacher evaluation scores than their non-SWD classmates (-0.32 and -0.158 standard deviation gaps respectively). We also note that the teacher evaluation score quality gap in the highest FRL bin is reliably negative and significant for all disability subgroups (ranging from -0.100 to -0.321 standard deviations). Additionally, there is one teacher quality measure that is consistently significant across all disability subgroups with SETs and FRL bins—novice teachers. The novice gap ranges from 5.8% (for SLDs in the lowest FRL bin) to 16.6% (for students with autism in the highest FRL bin), suggesting that SETs are consistently less experienced than the average math teacher for students without disabilities. Furthermore, across all disability subgroups, the novice TQG is significantly larger in the highest FRL schools compared to their peers in the lowest FRL bin.

Discussion & Policy Implications

In this study, we provide some of the first evidence documenting the extent of TQGs, not only between students with and without disabilities, but also differences in these gaps in more or less advantaged schools. We find that, overall, students with disabilities (SWD) learning in general education classrooms are more likely to experience teachers with lower VAMs relative to their non-SWD peers. Although, consistent with previous studies, we find that, on average, students at more disadvantaged schools in Los Angeles have lower quality teachers in terms of VAMs, teacher evaluation scores, and hiring scores, we nonetheless show that teacher quality gaps between SWDs and non-SWDs in general education classrooms does not increase with school-level disadvantage.

We find some differences by school subject, though we many of our point estimates are similar across subjects. For math, we do not find any TQGs according to teacher quality measures that are more easily observable to principals: teacher evaluation scores and novice teacher status. This suggests that principals are not actively sorting students with disabilities into classrooms with perceivably worse teacher characteristics and so there may be more fundamental, unobserved factors driving these patterns. For ELA teachers, we find significant differences for teacher evaluation scores and likelihood to have a novice teacher, though no significant differences across VAMs. These findings suggest that while there may be some sorting across observables for ELA students, these do not result in exposure to lower quality ELA teachers as measured by VAMs.

In our subgroup analysis across both subjects, we find evidence that TQGs are concentrated within students with specific learning disabilities. Students with autism or speech/language impairment do not seem to be placed in classrooms with teachers who are different from teachers of the average non-SWD student.

The new evidence we provide on SWD quality gaps contributes to a growing literature addressing the contexts and needs of both students with disabilities and the educators who teach them. Our finding of math VAM quality gaps between students with and without disabilities across all FRL bins suggests that schools, districts, and states should be cognizant of the ways in which they distribute teachers, particularly if schools are trying to adhere to the *Endrew F. v. Douglas County School District* decision to ensure equitable outcomes for SWDs. Existing literature shows that our case is not unique; schools tend to assign novice or less-effective teachers to larger proportions of low-performing students (e.g. Bruno et al, 2019; Kalogrides et

al, 2013; Lankford et al, 2002). Additionally, our variance decomposition suggests that the majority of SWD TQGs occur due to within school factors rather than between school factors.

For practitioners, the implication is that solutions to the SWD quality gaps does not necessarily have to come from district and state policies aimed at recruiting and retaining higher quality teachers overall—though these avenues can certainly help schools obtain more high-quality teachers. Instead, our estimates suggest that a more immediate solution could be to shift student compositions amongst existing teachers within a school. One potential avenue to accomplish this is to structure salary schedules so that teachers who teach SWDs receive salary enhancements, encouraging high quality teachers to take on these potentially more challenging teaching environments. While such a salary scheme would need to be negotiated into districts' collective bargaining agreements, states can also provide salary enhancements regardless of district policies.

Our findings suggest several avenues for future research, particularly qualitative work. Future qualitative interviews and observations could explore what, if any, factors principals take into consideration while creating each classroom. It is probable that teacher characteristics beyond those in the current study are used to determine how students are matched to teachers. For example, principals may pair certain SWDs with a teacher who is particularly strong at consistently engaging their students in classroom activities or who have strong classroom management skills. Furthermore, these traits may play an important role for improving SWD's academic outcomes. Similarly, future work could explore whether some general education teachers are empirically better at improving outcomes for SWD compared to other teachers. Researchers could then use both quantitative and qualitative data to determine which characteristics are strongly correlated with these outcomes and learn more about their teaching

practices. This empirical work could help practitioners move towards the end goal of more equitable academic outcomes for SWD.

Endnotes

1. For example, we do not include students who attend home or hospital schools, special education centers, nor community day schools.
2. While special education teachers are evaluated on the same observation instrument and hiring criteria as general education teachers, researchers and practitioners argue that these shared measures should not be used to measure special education teacher quality since special education teachers' work responsibilities and preparation programs are different from those for general education teachers (see Brownell, Ross, Colon, & McCallum, 2005 for a review). Additionally, recent research on teacher evaluations suggest that SETs may systematically receive lower evaluation scores since effective teaching looks different for SETs than GETs--particularly given the individualized nature of special education (Johnson and Semmelroth, 2013; Liu et al., 2019; Jones and Brownell, 2013).
3. For completeness, we provide the results in Appendix Tables 3 and 4. Given these concerns we are hesitant to say whether these results are indicative of the existence or non-existence of quality gaps among SWDs with special education teachers.
4. While typically IDEA only requires districts to designate a primary disability, along with blindness and deafness as secondary disabilities, LAUSD operates under a consent decree that requires more detailed tracking (Weintraub, Myers, Hehir, Jaque-Anton, 2008). For our main analysis, we are focused on whether students have any disabilities listed. Consequently, we do not separately account for students with multiple disabilities. In disability-specific analysis, we include any students who have that disability subcategory listed in their IEP (including those with multiple disabilities). We have also run analysis

that excludes students with multiple disabilities and find little difference. Results are available upon request.

5. In analysis not shown, we instead disaggregate schools by share of students who are underrepresented minorities or have low prior test scores. The results are qualitatively similar and available upon request.
6. As a sensitivity check, we have also split schools into four bins (<70, 70-<95, 95-<0.978, >=0.978), and five bins (<70, 70-<80, 80-<90, 90-<95, >=95). Results are similar to those found in our main tables and available upon request.
7. Specifically, we include four types of testing accommodation flags: technology (i.e. text-to-speech software), setting (i.e. small group setting), time (i.e. extended time), and format (i.e. streamlined version of text).
8. Only grades 3-8 have test scores that are usable in standard Value Added Measures, so we calculate VAMs only for students in grades 4-8 (leaving out grade 3 to ensure there is a lagged score). The teacher quality literature has used multiple different ways to measure teacher value-added. As a robustness check, we also estimate one-year teacher value-added measures that use teacher fixed effects and includes student- and classroom-level demographics (see Appendix B for more details about the construction of these models). To address concerns that student characteristics are endogenous to teacher value-added in time *t*, we use teacher's value-added score in t-1 as a measure for teacher quality in time *t*. Additionally, we create an alternative VAM score for teachers that exclude students with disabilities from VAM calculations. These results are presented in Appendix Table 5 and similar to the ones we show in our main tables.

9. Teacher evaluation may be deferred for employees with ten or more years of satisfactory service, have not received a "notice of unsatisfactory act of service" in the past four years, and had fewer than 13 unprotected absences in the past year.
10. We also analyzed results by teachers' final evaluation score, which only has three values: below standard performance, meets standard performance, and exceeds standard performance. Since less than 5% of teachers each year do not pass the evaluation, we focus our main results on the average score across all subcomponents. The average score across all subcomponents does not necessarily map onto the final evaluation score (though it very rarely does not match) and has the additional benefit of having more variation to distinguish between teacher scores. Results for final evaluation scores available upon request.
11. We have also analyzed a few alternative measures for teacher evaluation scores. Following Kraft et al (2018), we create a measure of overall performance using a graded response model for all subcomponents, *theta*, as well as a residualized *theta* measure that removes classroom- and school-level student demographic variation. However, since teachers are not all assessed on the same components, we also create a *theta* based only on the three subcomponents that are mandatory for all teachers, as well as a residualized *theta* score based on these three subcomponents. Finally, we also individually analyze the raw scores for each mandatory subcomponent. Across all these differing teacher evaluation measures, we find little evidence of SWD vs non-SWD teacher quality gaps. All results are displayed in Appendix Tables 6 and 7.
12. As a sensitivity check, we also define "novice teacher" as those with five or fewer years of experience. Results are qualitatively similar.

13. We do not present results for the “other disabilities” subgroup since this group encompasses a large range of disabilities from emotional disturbance to intellectual disability and interpreting any potential gaps would be difficult. However, for completeness, we include this indicator variable in our VAM calculations.

Citations

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95-135.

Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, 27(5), 483–503.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the Short Careers of High Achieving Teachers in Schools with Low-Performing Students. *American Economic Review*, 95(2), 166-171.

Blanton, L. P., Pugach, M. C., & Boveda, M. (2018). Interrogating the intersections between general and special education in the history of teacher education reform. *Journal of Teacher Education*, *69*(4), 354-366.

Blanton, L. P., Pugach, M. C., & Florian, L. (2011). Preparing general educators to improve outcomes for students with disabilities. *Washington, DC: American Association of Colleges of Teacher Education and National Council for Learning Disabilities*.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). Consequences? The Impact of Assessment and Accountability on Teacher Recruitment and Retention Are There Unintended Consequences? *Public Finance Review*, 36, 88–111. https://doi.org/10.1177/1091142106293446

Brownell, M. T., Ross, D. D., Colón, E. P., & McCallum, C. L. (2005). Critical features of special education teacher preparation: A comparison with general teacher education. *The Journal of Special Education*, *38*(4), 242-252.

Bruno, P. and Strunk, K. O. (2019). Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/0162373719865561

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

Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, 30(3), 449–465.

Chudowsky, N., Chudowsky, V., and Keber, N. (2009). State Test Score Trends Through 2007–08, Part 4: Has Progress Been Made in Raising Achievement for Students with Disabilities? Washington, DC: Center on Education Policy.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, 24, 377–392. https://doi.org/10.1016/j.econedurev.2004.06.008

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6), 673–682.

Endrew F. v. Douglas County School District RE–1, 580 U.S. ___ (2017)

Gilmour, A. F., & Henry, G. T. (2018a). Who Are the Classmates of Students With Disabilities in Elementary Mathematics Classrooms?. *Remedial and Special Education*, 41(1), 18-27.

Gilmour, A., & Henry, G. (2018). A comparison of teacher quality in math for late elementary and middle school students with and without disabilities. *The Elementary School Journal*, 118(3), 426–451.

Gilmour, A. F., & Wehby, J. H. (2019). The association between teaching students with disabilities and teacher turnover. Accepted in *Journal of Educational Psychology*. doi:10.1037/edu0000

Goldhaber, D., & Brewer, D. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis*, 22(2), 129–145. https://doi.org/10.3102/01623737022002129

Goldhaber, D., & Brewer, D. (2001). Evaluating the Evidence on Teacher Certification: A Rejoinder. *Educational Evaluation and Policy Analysis*, 23(1), 79–86.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, 44(5), 293–307.

Goldhaber, D., Quince, V., Theobald, R. (2018). How Did It Get This Way? Disentangling the Sources of Teacher Quality Gaps Across Two States. CALDER Working Paper No. 209-1118-1

Hanushek, E. A., Kane, T.J., & Rivkin, S.G. (2004). Why Public Schools Lose Teachers. *The Journal of Human Resources*, 39(2), 362-354.

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels. *Statistics and Public Policy*, 3(1), 1–10. https://doi.org/10.1080/2330443X.2016.1182878

Johnson, E., & Semmelroth, C. L. (2014). Special Education Teacher Evaluation. *Assessment for Effective Intervention*, 39(2), 71–82. https://doi.org/10.1177/1534508413513315

Jones, N. D. & Brownell, M. (2013a). Examining the Use of Classroom Observations in the Evaluation of Special Education Teachers. *Assessment for Effective Intervention,* 39, 112-124.

Jones, N. D., Buzick, H., and Turkan, S. (2013b). Including Students With Disabilities and English Learners in Measures of Educator Effectiveness. *Educational Researcher*, 42, 234-241.

Kalogrides, D., & Loeb, S. (2013). Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools. *Educational Researcher*, 42(6), 304–316. https://doi.org/10.3102/0013189X13495087

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6), 615-631.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47, 180–195. https://doi.org/10.1016/J.ECONEDUREV.2015.01.006

Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, 12(2), 241-279.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Education Evaluation and Policy Analysis*, 24(1), 37-62.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, 13(2), 125–145.

National Center for Education Statistics. (2015). The Condition of Education. Retrieved from https://nces.ed.gov/pubs2015/2015144.pdf

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130, 105-119.

Rice, J. K. (2010). The Impact of Teacher Experience: Examining the Evidence and Policy Implications. Brief No. 11. National center for analysis of longitudinal data in education research.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247-252.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, 72(2-3), 104-122.

Schulte, A. C., & Stevens, J. J. (2015). Once, Sometimes, or Always in Special Education Mathematics Growth and Achievement Gaps. *Exceptional Children*, 81, 370-387.

Schulte, A. C., Elliott, J. J., Tindal, S. N., & Nese, G. (2016). Achievement Gaps for Students with Disabilities: Stable, Widening, or Narrowing on a State-wide Reading Comprehension Test? *Journal of Educational Psychology*, 16, 925–942. https://doi.org/10.1037/edu0000107

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3), 97-118.

Swaak, Taylor. (2020, January 22). For the first time in more than 20 years, LAUSD is in full control of its special ed system. As parents worry about accountability, the district shifts its focus. *The New York Times*, Retrieved from http://laschoolreport.com/for-the-first-time-in-20-years-lausd-is-in-full-control-of-its-special-ed-system-as-parents-worry-about-accountability-the-district-shifts-its-focus/

U.S. Department of Education. (2019). The NCES Fast Facts Tool provides quick answers to many education questions (National Center for Education Statistics).
Weintraub, F. J., Myers, R. M., Hehir, T., & Jaque-Anton, D. (2008). A Contextual Overview of the Modified Consent Decree in the Los Angeles Unified School District. *Journal of Special Education Leadership*, *21*(2), 51-57.

Tables

Table 1. Teacher Characteristics in Math Classes for FRL School Bins

	Panel A. VAM Sample				Panel B. Teacher Eval Sample				Panel C. Hiring Score Sample				Panel D. Novice Sample			
FRL Group	**<=70%**	**70-95%**	**>=95%**	**Overall**	**<=70%**	**70-95%**	**>=95%**	**Overall**	**<=70%**	**70-95%**	**>=95%**	**Overall**	**<=70%**	**70-95%**	**>=95%**	**Overall**
Schools	121	133	341	595	121	128	351	600	106	98	284	488	123	134	362	619
General Education Teacher Characteristics																
Teachers	836	993	3,121	4,950	983	1,214	4,094	6,291	257	217	785	1,259	2,423	2,477	8,207	13,107
Teacher-Years	2,764	3,398	10,253	16,415	1,783	2,338	7,749	11,870	593	512	1,783	2,888	7,519	8,087	26,226	41,832
Mean Exp	9.41	9.27	9.31	9.32	8.82	8.92	9.11	9.02	4.65	4.81	5.14	4.98	9.36	9.36	9.37	9.36
%Novice	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.34	0.31	0.29	0.31	0.03	0.03	0.03	0.03
%MAH	0.34	0.40	0.39	0.38	0.34	0.37	0.36	0.36	0.39	0.39	0.36	0.37	0.34	0.36	0.36	0.36
%Fem	0.72	0.68	0.66	0.68	0.78	0.76	0.73	0.74	0.90	0.80	0.85	0.85	0.80	0.75	0.75	0.76
%White	0.53	0.36	0.25	0.33	0.52	0.31	0.22	0.29	0.62	0.28	0.25	0.33	0.53	0.35	0.23	0.31
%Black	0.06	0.10	0.10	0.09	0.06	0.09	0.09	0.08	0.02	0.05	0.09	0.07	0.06	0.10	0.09	0.08
%Hispanic	0.19	0.32	0.53	0.42	0.22	0.40	0.58	0.48	0.19	0.41	0.51	0.42	0.20	0.36	0.56	0.45
%Have VAM	1.00	1.00	1.00	1.00	0.61	0.61	0.56	0.58	0.46	0.63	0.57	0.56	0.53	0.58	0.51	0.52
Mean VAM	-0.02	-0.14	-0.05	-0.07	-0.11	-0.19	-0.13	-0.14	-0.04	-0.20	-0.08	-0.10	-0.02	-0.14	-0.05	-0.07
%Have Teacher Eva	0.29	0.30	0.33	0.32	1.00	1.00	1.00	1.00	0.57	0.61	0.57	0.58	0.25	0.29	0.30	0.29
Mean Teacher Eval	0.35	0.14	0.11	0.16	0.36	0.19	0.16	0.20	0.28	0.13	0.16	0.18	0.36	0.19	0.16	0.20
%Have Hiring Score	0.06	0.07	0.08	0.08	0.17	0.15	0.13	0.14	1.00	1.00	1.00	1.00	0.07	0.07	0.07	0.07
Mean Hiring Score	0.10	0.08	-0.08	-0.01	0.18	0.06	-0.03	0.03	0.13	0.06	-0.04	0.01	0.13	0.06	-0.04	0.01

Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. Each column represents teacher characteristics in schools binned by the percent of FRL eligible students. A school's FRL bin is defined by taking a three year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. MAH represents any teacher with a master's (or higher) degree. VAM, Teacher Eval, and Hiring Score are z-scored measures for value-added, evaluation scores, and hiring scores, respectively. %Have indicates what percent of the given sample has a VAM, Teacher Eval, or Hiring Score measure. Panels A, B, C represent the VAM, Teacher Eval, and Hiring Score samples, repsectively. Panel D represents the population sample.

Table 2. Average Math Teacher Quality, by FRL Bin

	(1) VAM	(2) Teach Eval	(3) Hiring Score	(4) Novice
A. <70% FRL	-0.016	0.361	0.123	0.034
Std. Dev	(0.529)	(0.729)	(1.084)	(0.180)
n	118,370	55,682	16,546	223,871
B. 70% - <95% FRL	-0.135	0.19	0.061	0.03
Std. Dev	(0.620)	(0.826)	(0.710)	(0.172)
n	141,686	69,980	16,777	245,351
Disadv. Gap [p-value]	[0.023]	[0.01]	[0.48]	[0.85]
C. >= 95% FRL	-0.055	0.161	-0.039	0.031
Std. Dev	(0.617)	(0.854)	(0.928)	(0.173)
n	357,985	213,663	49,676	708,089
Disadv. Gap [p-value]	[0.289]	[0]	[0.114]	[0.754]

Exposure rates calculated from Eq (1). Standard errors clustered at school level. Disadv. Gap represents how similar the FRL bins from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Table 3. SWD vs non-SWD Math Teacher Quality Gaps by Teacher Type

	(1) VAM	(2) Teach Eval	(3) Hiring Score	(4) Novice
SWD	-0.088	0.174	-0.002	0.033
Non-SWD	-0.064	0.202	0.015	0.031
Gap	-0.024***	-0.028*	-0.017	0.002
Std Error	(0.007)	(0.011)	(0.021)	(0.001)
SWD n	50,521	27,276	6,706	91,062
non-SWD n	567,520	312,049	76,293	1,086,249
between variance	*0.253*	*0.341*	*0.475*	*0.332*
within variance	*0.747*	*0.659*	*0.525*	*0.668*

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Table 4. Teacher Quality Gaps for Math classes by FRL Bins

SWD with GET vs non-SWD				
	(1) VAM	(2) Teach Eval	(3) Hiring Score	(4) Novice
A. <70% FRL				
SWD	-0.059	0.335	0.105	0.034
Non-SWD	-0.013	0.363	0.125	0.034
Gap	-0.047***	-0.028	-0.019	0
Std Error	(0.014)	(0.026)	(0.050)	(0.003)
SWD n	9,258	4,397	1,266	16,865
non-SWD n	109,112	51,285	15,280	207,006
between variance	*0.26*	*0.378*	*0.427*	*0.041*
within variance	*0.74*	*0.622*	*0.573*	*0.959*
B. 70% - <95% FRL				
SWD	-0.148	0.153	0.061	0.037
Non-SWD	-0.134	0.194	0.061	0.03
Gap	-0.014	-0.041	-0.001	0.007
Std Error	(0.018)	(0.024)	(0.035)	(0.004)
SWD n	12,152	5,913	1,429	20,005
non-SWD n	129,534	64,067	15,348	225,346
between variance	*0.281*	*0.424*	*0.61*	*0.357*
within variance	*0.719*	*0.576*	*0.39*	*0.643*
Disadv. Gap [p-value]	[0.143]	[0.71]	[0.757]	[0.19]
C. >= 95% FRL				
SWD	-0.073	0.139	-0.058	0.032
Non-SWD	-0.053	0.162	-0.037	0.031
Gap	-0.019*	-0.023	-0.021	0.001
Std Error	(0.008)	(0.014)	(0.029)	(0.001)
SWD n	29,111	16,966	4,011	54,192
non-SWD n	328,874	196,697	45,665	653,897
between variance	*0.24*	*0.298*	*0.453*	*0.393*
within variance	*0.76*	*0.702*	*0.547*	*0.607*
Disadv. Gap [p-value]	[0.083]	[0.879]	[0.974]	[0.865]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Table 5. Math Teacher Quality Gaps by Disability Type (vs. No Disability) and FRL Bins

	SWD with GET vs non-SWD			
	(1)	(2)	(3)	(4)
Disability & FRL Group	VAM	Teach Eval	Hiring Score	Novice
A. Specific Learning				
<70%	-0.074***	-0.049	-0.062	0.002
Std Error	(0.019)	(0.047)	(0.075)	(0.004)
SWD n	5,071	2,110	560	7,903
non-SWD n	109,112	51,285	15,280	207,006
70-<95%	-0.022	-0.087*	-0.02	0.01
Std Error	(0.019)	(0.038)	(0.044)	(0.006)
SWD n	8,078	3,429	859	11,935
non-SWD n	129,534	64,067	15,348	225,346
Disadv. Gap [p-value]	[0.051]	[0.519]	[0.627]	[0.252]
>=95%	-0.029**	-0.053**	-0.016	0.004
Std Error	(0.009)	(0.020)	(0.040)	(0.002)
SWD n	19,997	9,930	2,446	31,959
non-SWD n	328,874	196,697	45,665	653,897
Disadv. Gap [p-value]	[0.032]	[0.928]	[0.592]	[0.661]
B. Autism				
<70%	0	0.012	-0.006	-0.004
Std Error	(0.021)	(0.035)	(0.081)	(0.004)
SWD n	1,405	715	199	2,834
non-SWD n	109,112	51,285	15,280	207,006
70-<95%	0.01	0.063	0.053	0.004
Std Error	(0.032)	(0.041)	(0.047)	(0.005)
SWD n	1,126	641	139	2,090
non-SWD n	129,534	64,067	15,348	225,346
Disadv. Gap [p-value]	[0.784]	[0.338]	[0.529]	[0.226]
>=95%	0.005	0.033	-0.034	-0.003
Std Error	(0.016)	(0.026)	(0.056)	(0.003)
SWD n	2,267	1,405	325	4,632
non-SWD n	328,874	196,697	45,665	653,897
Disadv. Gap [p-value]	[0.851]	[0.619]	[0.779]	[0.853]
C. Speech/Language				
<70%	0.022	-0.021	0.153	-0.004
Std Error	(0.031)	(0.039)	(0.089)	(0.004)
SWD n	888	821	260	3,098
non-SWD n	109,112	51,285	15,280	207,006
70-<95%	0.076*	0.056	-0.03	-0.002
Std Error	(0.029)	(0.039)	(0.070)	(0.004)
SWD n	1,223	1,137	231	3,647
non-SWD n	129,534	64,067	15,348	225,346
Disadv. Gap [p-value]	[0.207]	[0.158]	[0.107]	[0.752]
>=95%	0.058***	0.076***	0.002	-0.006**
Std Error	(0.014)	(0.023)	(0.042)	(0.002)
SWD n	3,821	4,249	855	13,701
non-SWD n	328,874	196,697	45,665	653,897
Disadv. Gap [p-value]	[0.293]	[0.03]	[0.125]	[0.706]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Appendix Table 1. Teacher Quality Gaps for ELA classes by FRL Bins

SWD with GET vs non-SWD				
	(1) VAM	(2) Teach Eval	(3) Hiring Score	(4) Novice
A. <70% FRL				
SWD	-0.119	0.244	0.261	0.033
Non-SWD	-0.12	0.329	0.229	0.032
Gap	0.001	-0.085**	0.032	0.001
Std Error	(0.024)	(0.028)	(0.042)	(0.003)
SWD n	9,839	4,318	1,467	17,366
non-SWD n	121,328	52,275	16,386	218,257
between variance	*0.439*	*0.393*	*0.51*	*0.449*
within variance	*0.561*	*0.607*	*0.49*	*0.551*
B. 70% - <95% FRL				
SWD	-0.378	0.136	0.236	0.039
Non-SWD	-0.366	0.208	0.152	0.031
Gap	-0.012	-0.072*	0.084	0.008*
Std Error	(0.037)	(0.030)	(0.068)	(0.004)
SWD n	13,112	6,294	1,915	20,751
non-SWD n	135,557	66,028	17,872	229,389
between variance	*0.508*	*0.358*	*0.553*	*0.264*
within variance	*0.492*	*0.642*	*0.447*	*0.736*
Disadv. Gap [p-value]	[0.76]	[0.752]	[0.521]	[0.163]
C. >= 95% FRL				
SWD	-0.255	0.155	-0.012	0.042
Non-SWD	-0.228	0.174	-0.01	0.035
Gap	-0.027	-0.019	-0.002	0.007***
Std Error	(0.020)	(0.015)	(0.028)	(0.002)
SWD n	30,653	17,789	5,558	55,320
non-SWD n	340,481	199,019	54,715	660,545
between variance	*0.451*	*0.317*	*0.39*	*0.448*
within variance	*0.549*	*0.683*	*0.61*	*0.552*
Disadv. Gap [p-value]	[0.368]	[0.037]	[0.487]	[0.111]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Apx Table 2. Teacher Quality Gaps for Math classes by FRL Bins

	SWD with SET vs non-SWD		
	(1)	(2)	(3)
	Teach Eval	Hiring Score	Novice
A. <70% FRL			
SWD	0.16	-0.138	0.119
Non-SWD	0.363	0.125	0.033
Gap	-0.203	-0.263	0.086***
Std Error	(0.119)	(0.152)	(0.018)
SWD n	1,468	1,089	5,933
non-SWD n	51,285	15,280	207,006
between variance	*0.365*	*0.401*	*0.037*
within variance	*0.635*	*0.599*	*0.963*
B. 70% - <95% FRL			
SWD	0.011	-0.166	0.133
Non-SWD	0.194	0.061	0.03
Gap	-0.183*	-0.228*	0.103***
Std Error	(0.077)	(0.115)	(0.016)
SWD n	4,118	2,639	11,717
non-SWD n	64,067	15,348	225,346
between variance	*0.395*	*0.512*	*0.245*
within variance	*0.605*	*0.488*	*0.755*
Disadv. Gap [p-value]	[0.885]	[0.853]	[0.478]
C. >= 95% FRL			
SWD	0.047	-0.155	0.164
Non-SWD	0.162	-0.037	0.031
Gap	-0.115*	-0.118	0.134***
Std Error	(0.048)	(0.074)	(0.009)
SWD n	13,017	9,334	33,795
non-SWD n	196,697	45,665	653,897
between variance	*0.287*	*0.423*	*0.349*
within variance	*0.713*	*0.577*	*0.651*
Disadv. Gap [p-value]	[0.493]	[0.392]	[0.019]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Apx Table 3. Teacher Quality Gaps for ELA classes by Disability Type (vs. No Disability) and FRL Bins

	SWD with GET vs non-SWD			
	(1)	(2)	(3)	(4)
Disability & FRL Group	VAM	Teach Eval	Hiring Score	Novice
A. Specific Learning				
<70%	-0.033	-0.123***	0.055	0
Std Error	(0.031)	(0.034)	(0.066)	(0.004)
SWD n	5,131	1,942	630	7,901
non-SWD n	121,328	52,275	16,386	218,257
70-<95%	-0.022	-0.103**	0.132	0.006
Std Error	(0.043)	(0.038)	(0.086)	(0.005)
SWD n	8,231	3,505	1,032	11,902
non-SWD n	135,557	66,028	17,872	229,389
Disadv. Gap [p-value]	[0.833]	[0.695]	[0.478]	[0.412]
>=95%	-0.043*	-0.039*	0.031	0.010***
Std Error	(0.022)	(0.019)	(0.038)	(0.003)
SWD n	20,269	10,141	3,283	31,879
non-SWD n	340,481	199,019	54,715	660,545
Disadv. Gap [p-value]	[0.785]	[0.03]	[0.756]	[0.044]
B. Autism				
<70%	0.069	-0.009	0.049	-0.001
Std Error	(0.042)	(0.039)	(0.075)	(0.004)
SWD n	1,485	690	210	2,905
non-SWD n	121,328	52,275	16,386	218,257
70-<95%	0.015	-0.011	0.099	0.005
Std Error	(0.053)	(0.051)	(0.104)	(0.006)
SWD n	1,165	632	171	2,106
non-SWD n	135,557	66,028	17,872	229,389
Disadv. Gap [p-value]	[0.42]	[0.979]	[0.701]	[0.383]
>=95%	0.071	0.03	0.03	0
Std Error	(0.045)	(0.027)	(0.054)	(0.003)
SWD n	2,273	1,435	418	4,613
non-SWD n	340,481	199,019	54,715	660,545
Disadv. Gap [p-value]	[0.974]	[0.402]	[0.831]	[0.852]
C. Speech/Language				
<70%	0.207***	-0.013	0.012	0.001
Std Error	(0.058)	(0.042)	(0.084)	(0.005)
SWD n	936	814	274	3,134
non-SWD n	121,328	52,275	16,386	218,257
70-<95%	0.308***	0.05	-0.025	-0.003
Std Error	(0.049)	(0.040)	(0.086)	(0.004)
SWD n	1,242	1,134	233	3,652
non-SWD n	135,557	66,028	17,872	229,389
Disadv. Gap [p-value]	[0.182]	[0.277]	[0.758]	[0.548]
>=95%	0.230***	0.058**	-0.016	-0.008***
Std Error	(0.033)	(0.021)	(0.035)	(0.002)
SWD n	3,887	4,296	907	13,749
non-SWD n	340,481	199,019	54,715	660,545
Disadv. Gap [p-value]	[0.732]	[0.134]	[0.76]	[0.099]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Apx Table 4. Math Teacher Quality Gaps by Disability Type (vs. No Disability) and FRL Bins

	SWD with SET vs non-SWD		
	(1)	(2)	(3)
Disability & FRL Group	Teach Eval	Hiring Score	Novice
A. Specific Learning			
<70%	-0.408	-0.367	0.058*
Std Error	(0.288)	(0.197)	(0.025)
SWD n	429	314	2,109
non-SWD n	51,285	15,280	207,006
70-<95%	-0.321**	-0.255	0.100***
Std Error	(0.096)	(0.161)	(0.024)
SWD n	1,876	1,256	5,481
non-SWD n	64,067	15,348	225,346
Disadv. Gap [p-value]	[0.775]	[0.662]	[0.228]
>=95%	-0.158*	-0.17	0.119***
Std Error	(0.070)	(0.101)	(0.012)
SWD n	6,435	4,324	17,008
non-SWD n	196,697	45,665	653,897
Disadv. Gap [p-value]	[0.398]	[0.375]	[0.03]
B. Autism			
<70%	-0.068	-0.168	0.100***
Std Error	(0.094)	(0.173)	(0.022)
SWD n	730	567	2,510
non-SWD n	51,285	15,280	207,006
70-<95%	-0.166	-0.166	0.122***
Std Error	(0.088)	(0.128)	(0.018)
SWD n	1,260	801	3,455
non-SWD n	64,067	15,348	225,346
Disadv. Gap [p-value]	[0.448]	[0.991]	[0.442]
>=95%	-0.100*	-0.071	0.166***
Std Error	(0.044)	(0.069)	(0.012)
Std Error	3,955	3,239	10,075
SWD n	196,697	45,665	653,897
non-SWD n	[0.762]	[0.599]	[0.008]
C. Speech/Language			
<70%	-0.219	-0.252	0.057
Std Error	(0.161)	(0.275)	(0.030)
SWD n	50	34	154
non-SWD n	51,285	15,280	207,006
70-<95%	-0.16	-0.205	0.059**
Std Error	(0.097)	(0.144)	(0.019)
SWD n	155	90	396
non-SWD n	64,067	15,348	225,346
Disadv. Gap [p-value]	[0.752]	[0.88]	[0.973]
>=95%	-0.139*	-0.028	0.136***
Std Error	(0.058)	(0.082)	(0.014)
SWD n	724	437	1,498
non-SWD n	196,697	45,665	653,897
Disadv. Gap [p-value]	[0.638]	[0.434]	[0.018]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Apx Table 5. Math VAM Teacher Quality Gaps by VAM Approach and FRL Bins

SWD with GET vs non-SWD	(1) One-step VAM	(2) Two-step VAM
A. <70% FRL		
SWD	0.275	-0.093
Non-SWD	0.411	-0.023
Gap	-0.136***	-0.071***
Std Error	(0.029)	(0.019)
SWD n	6,090	6,090
non-SWD n	72,841	72,841
between variance	*0.255*	*0.339*
within variance	*0.745*	*0.661*
B. 70% - <95% FRL		
SWD	-0.145	-0.2
Non-SWD	-0.041	-0.183
Gap	-0.104***	-0.017
Std Error	(0.025)	(0.025)
SWD n	8,022	8,022
non-SWD n	85,772	85,772
between variance	*0.245*	*0.34*
within variance	*0.755*	*0.66*
Disadv. Gap [p-value]	[0.401]	[0.086]
C. >= 95% FRL		
SWD	-0.131	-0.101
Non-SWD	-0.052	-0.081
Gap	-0.079***	-0.02
Std Error	(0.013)	(0.011)
SWD n	18,846	18,846
non-SWD n	209,948	209,948
between variance	*0.206*	*0.314*
within variance	*0.794*	*0.686*
Disadv. Gap [p-value]	[0.073]	[0.023]

Exposure rates calculated from Eq (1). Between/Within variance calculated from Eq (1) with the addition of school fixed-effects. Standard errors clustered at school level. Disadv. Gap represents how similar the Free-Reduced Lunch bins are from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. Exp represents years of experience and is top-coded at 10 years. Novice is defined by any teacher with fewer than 2 years of experience. VAM (value-added measure), Teacher Eval, and Hiring Score are z-scored measures.

Appendix Table 6. Different Teacher Evaluation Score Measures by School FRL Bins

	(1) Previous Eval	(2) Resid Previous Eval	(3) Theta (all, previous eval)	(4) Resid Theta (all, previous eval)	(5) Theta (main 3, previous eval)	(6) Resid Theta (main 3, previous eval)	(7) Discussion Techniques and Student Participation (3b2)	(8) Standards-Based Projects, Activities, Assignments (3c1)	(9) Student Feedback (3d3)
A. <70% FRL	0.355	0.012	0.29	0.075	0.284	-0.013	2.795	2.897	2.85
Std. Dev	(0.733)	(0.941)	(0.807)	(0.801)	(0.758)	(0.959)	(0.414)	(0.320)	(0.361)
n	57,150	56,277	56,913	56,361	57,133	56,277	57,133	57,113	57,133
B. 70% - <95% FRL	0.18	0.072	0.084	0.024	0.096	0.019	2.724	2.817	2.778
Std. Dev	(0.829)	(0.960)	(0.904)	(0.908)	(0.897)	(0.992)	(0.463)	(0.413)	(0.448)
n	74,068	72,527	73,826	72,544	74,051	72,527	73,966	74,008	73,989
Disadv. Gap [p-value]	[0.01]	[0.44]	[0.008]	[0.497]	[0.004]	[0.588]	[0.03]	[0.001]	[0.027]
C. >= 95% FRL	0.154	0.035	0.073	0.08	0.08	0.033	2.73	2.812	2.753
Std. Dev	(0.855)	(0.959)	(0.918)	(0.929)	(0.912)	(0.990)	(0.475)	(0.418)	(0.460)
n	226,652	218,235	225,872	219,065	226,498	218,235	226,337	226,460	226,227
Disadv. Gap [p-value]	[0]	[0.733]	[0]	[0.922]	[0]	[0.38]	[0.012]	[0]	[0]

Exposure rates calculated from Eq. (1). Between/Within variance calculated from Eq. (1) with the inclusion of school fixed-effects. Disadv. Gap represents how similar the FRL bins from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. *Previous Eval* is the teacher's most recent evaluation score, before the current year. *Resid Previous Eval* represents the residualized teacher evaluation score after accounting for classroom- and school-level demographics. *Theta (all)* is the theta score calculated with the graded response model across all evaluation subcomponents. *Resid Theta (all)* is the residualized theta score after accounting for classroom- and school-level demographics. *Theta (req. 3)* is the theta score calculated with the graded response model, but only with the three required subcomponents. *Resid Theta (req. 3)* is the residualized theta score from the three required subcomponents after accounting for classroom- and school-level demographics. Columns (1)-(6) are standardized by year while Columns (7)-(9) represent raw scores. Sample restricted to observations with valid responses for every outcome.

Appendix Table 7. Math Teach Eval Quality Gaps by Different Evaluation Measures and FRL Bins

	SWD with GET vs non-SWD								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Previous Eval	Resid Previous Eval	Theta (all)	Resid Theta (all)	Theta (req. 3)	Resid Theta (req. 3)	Discussion Techniques and Student Participation (3b2)	Standards-Based Projects, Activities, Assignments (3c1)	Student Feedback (3d3)
A. <70% FRL									
SWD	0.316	0.054	0.223	0.028	0.231	-0.003	2.771	2.884	2.83
Non-SWD	0.343	0.007	0.275	0.056	0.273	-0.014	2.788	2.894	2.853
Gap	-0.027	0.047	-0.052	-0.028	-0.042	0.011	-0.017	-0.01	-0.023
Std Error	(0.029)	(0.030)	(0.030)	(0.030)	(0.028)	(0.032)	(0.014)	(0.010)	(0.013)
SWD n	3,880	3,880	3,880	3,880	3,880	3,880	3,880	3,880	3,880
non-SWD n	45,269	45,269	45,269	45,269	45,269	45,269	45,269	45,269	45,269
between variance	*0.391*	*0.289*	*0.429*	*0.427*	*0.389*	*0.242*	*0.335*	*0.349*	*0.349*
within variance	*0.609*	*0.711*	*0.571*	*0.573*	*0.611*	*0.758*	*0.665*	*0.651*	*0.651*
B. 70% - <95% FRL									
SWD	0.145	0.038	0.054	-0.005	0.067	-0.004	2.72	2.794	2.77
Non-SWD	0.183	0.072	0.081	0.005	0.098	0.014	2.732	2.813	2.781
Gap	-0.038	-0.034	-0.027	-0.01	-0.031	-0.018	-0.012	-0.02	-0.01
Std Error	(0.029)	(0.036)	(0.028)	(0.027)	(0.028)	(0.026)	(0.013)	(0.015)	(0.015)
SWD n	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000
non-SWD n	55,108	55,108	55,108	55,108	55,108	55,108	55,108	55,108	55,108
between variance	*0.463*	*0.325*	*0.499*	*0.475*	*0.404*	*0.207*	*0.368*	*0.353*	*0.357*
within variance	*0.537*	*0.675*	*0.501*	*0.525*	*0.596*	*0.793*	*0.632*	*0.647*	*0.643*
Disadv. Gap [p-value]	[0.782]	[0.087]	[0.544]	[0.657]	[0.791]	[0.49]	[0.799]	[0.577]	[0.505]
C. >= 95% FRL									
SWD	0.103	0.034	0.014	0.019	0.048	0.068	2.714	2.797	2.758
Non-SWD	0.126	0.027	0.046	0.048	0.062	0.039	2.724	2.805	2.749
Gap	-0.023	0.007	-0.031	-0.029	-0.013	0.029	-0.010	-0.007	0.008
Std Error	(0.016)	(0.015)	(0.017)	(0.017)	(0.017)	(0.016)	(0.010)	(0.006)	(0.006)
SWD n	14,466	14,466	14,466	14,466	14,466	14,466	14,466	14,466	14,466
non-SWD n	166,316	166,316	166,316	166,316	166,316	166,316	166,316	166,316	166,316
between variance	*0.326*	*0.234*	*0.317*	*0.315*	*0.282*	*0.216*	*0.238*	*0.259*	*0.241*
within variance	*0.674*	*0.766*	*0.683*	*0.685*	*0.718*	*0.784*	*0.762*	*0.741*	*0.759*
Disadv. Gap [p-value]	[0.902]	[0.236]	[0.548]	[0.978]	[0.386]	[0.601]	[0.691]	[0.851]	[0.031]

Exposure rates calculated from Eq. (1). Between/Within variance calculated from Eq. (1) with the inclusion of school fixed-effects. Disadv. Gap represents how similar the FRL bins from the least disadvantaged bin (<70% FRL). Observations are at student-teacher cell level pooled across school years 2014-15 to 2017-18. A school's FRL bin is defined by taking a four year average of the percent of FRL eligible students. *Previous Eval* is the teacher's most recent evaluation score, before the current year. *Resid Previous Eval* represents the residualized teacher evaluation score after accounting for classroom- and school-level demographics. *Theta (all)* is the theta score calculated with the graded response model across all evaluation subcomponents. *Resid Theta (all)* is the residualized theta score after accounting for classroom- and school-level demographics. *Theta (req. 3)* is the theta score calculated with the graded response model, but only with the three required subcomponents. *Resid Theta (req. 3)* is the residualized theta score from the three required subcomponents after accounting for classroom- and school-level demographics. Columns (1)-(6) are standardized by year while Columns (7)-(9) represent raw scores. Sample restricted to observations with valid responses for every outcome.

Appendix A. Teacher Assignment
(via Los Angeles Unified School District Human Resources: Staff Relations Handbook)

This information is intended to provide guidance to Principals so that they can comply with the LAUSD/UTLA Collective Bargaining Agreement and assure that teacher assignments best meet students' instructional needs and priorities.

Elementary School Assignments

1. In elementary schools, the LAUSD/UTLA Collective Bargaining Agreement (**CBA**) Article IX-A, Section 2.0 c (1) (ii) provides that the site administrator shall assign permanent teachers to **track** or **grade level** opening on the basis of seniority. Appropriate credential should be considered for Special Education assignments.

2. The Collective Bargaining Agreement does not provide teachers the right to select specific instructional programs, student performance levels or instructional clustering of students.

3. Principals can use preference forms (District's preferred method that will support an effective instructional program) or locally determined method to receive teachers' requests for assignments.

4. The site administrator **can and should make exceptions to the CBA provision** if he or she reasonably determines that the specific assignment is not in the best interest of the educational program.

5. Teachers with the specified credentials and required qualifications ("qualified") may request assignment to their grade level using a teacher preference form or other locally determined method. Submission of this form shall serve as a request for the assignment.

6. Administrators should review credentials, specific training, authorizations, performance indicators (i.e. pre/post assessment data, EL reclassification data, DIBELS) teacher status (Probationary 1 and 2) and evaluation/conduct records to inform their decision to assign a teacher to a specific class.

Secondary School Assignments

1. In secondary schools, Article IX-A, Section 2.0 d only provides teachers the right to a **department** selection on the basis of recent experience/seniority.

2. **Principals retain the authority** to assign teachers to particular classes and sections within a department.

3. Secondary principals must understand that the CBA does not confer the right for teachers to select either classes or "lines" on the master schedule.

4. Principals should take in consideration the best interest of the instructional program including specific training, authorizations, performance indicators (i.e. core subject end of the year assessments data, EL reclassification data, pre/post assessment data) teacher status (Probationary 1 and 2) and evaluation/conduct records to inform their decision to assign a teacher to a specific class.

5. Principals **can and should use objective data** as described above to assign teachers to classes.

6. Classes within a department shall be distributed by the Principal (or designee) in consultation with the **elected department chair**.

Appendix B. Two-step Average Residual and One-Year VAM Calculations

Two-step Average Residual VAM Calculations

We begin by using the following equation to create a residualized test score for student *i* in year *t*:

$$(1) \quad Ach_{ijst} = \delta_1 SameAch_{ijst-1} + \delta_2 SameAch^2_{ijst-1} + \delta_3 SameAch^3_{ijst-1} + \delta_4 OtherAch_{ijst-1} + \delta_5 OtherAch^2_{ijst-1} + \delta_6 OherAch^3_{ijst-1} + \boldsymbol{X}_{ijst}\boldsymbol{\theta} + \boldsymbol{T}_{jt}\boldsymbol{\Omega} + \varepsilon_{ijst}$$

where Ach_{ijst} is either math or ELA achievement, standardized within test and year, for student *i* with teacher *j* in school *s* and year *t*, $SameAch_{ijst-1}$ is the student's prior year score in the same subject, which enter as a cubic polynomial, $OtherAch_{ijst-1}$ is the student's prior year score in the other subject, which also enters as a cubic polynomial. In other words, if we are looking at math achievement, $SameAch_{ijst-1}$ would represent the student's prior math test score and $OtherAch_{ijst-1}$ would represent the student's prior ELA test score. X is a vector of student-, classroom- and grade-level demographics, and $\boldsymbol{T}_{jt}$ is a vector of teacher fixed effects. Specifically, X contains information about race, gender, free/reduced lunch status, English Language Learner status, student with disability status, and testing accommodation flags for technology (i.e. text-to-speech software), setting (i.e. small group setting), time (i.e. extended time), and format (i.e. streamlined version of text). Consequently, the residualized test scores are calculated in the following equation:

$$(2)\quad Ach^{*}_{ijst} = SameAch_{ijst} - \delta_1 SameAch_{ijst-1} - \delta_2 SameAch^2_{ijst-1} - \delta_3 SameAch^3_{ijst-1} - \delta_4 OtherAch_{ijst-1} - \delta_5 OtherAch^2_{ijst-1} - \delta_6 OherAch^3_{ijst-1} - \boldsymbol{X}_{ijst}\boldsymbol{\theta} = \boldsymbol{T}_{jt}\boldsymbol{\Omega} + \varepsilon_{ijst}$$

Next, students' residual scores in time *t* are averaged to create teacher value-added for teacher *j*, $\overline{A}_{jt}$. Residual average scores from prior years are used to calculate the best linear predictor of $\overline{A}_{jt}$ for teacher *j* in year *t* and forecasting coefficients, ψ, that minimizes the mean-squared error of the test score forecasts are selected:

$$(3)\quad \psi = \underset{\{\psi_1,\ldots,\psi_{t-1}\}}{\operatorname{argmin}} \sum_j \left(\overline{A}_{jt} - \sum_{s=1}^{t-1} \psi_s \overline{A}_{js}\right)^2$$

Finally, estimates of ψ from any year outside of *t* are used to calculate value-added for teacher *j* in year *t*.

<u>One-year VAM Calculations</u>

We estimate VAMs using the following model for each school year from 2012-2013 to 2016-17, and separately for each subject and level (elementary and secondary):

$$(4)\quad Ach_{ijst} = \beta_1 Ach^{math}_{ijst-1} + \beta_2 Ach^{ela}_{ijst-1} + \boldsymbol{X}_{ijst}\boldsymbol{\theta} + \boldsymbol{T}_{jt}\boldsymbol{\Omega} + \varepsilon_{ijst}$$

where Ach is either math or ELA achievement, standardized within test and year, for student *i* with teacher *j* in school *s* and year *t*. We control for students' achievement in the prior year in both math and ELA, since the inclusion of the second subject is helpful in mitigating bias due to sorting (Chetty, Friedman, & Rockoff, 2014) and to attenuate measurement error (Lockwood & McCaffrey, 2014). **X** is a vector of student demographic characteristics, including indicators of student race, gender, free- or reduced-price lunch eligibility, grade level, and English learner status. Johnson & Semmelroth (2013) argue that each disability type requires different teaching methods so we also include disability type indicators (Autism, Specific

Learning Disability, Speech/Language Impairment, and Other Disability) into the vector **X** in order to measure a teacher's overall effectiveness in achievement growth, as opposed to effectiveness towards specific disability types. The final component of **X** is an indicator which describes whether a student has testing accommodations. Specifically, we include four types of testing accommodation flags: technology (i.e. text-to-speech software), setting (i.e. small group setting), time (i.e. extended time), and format (i.e. streamlined version of text). Jones, Buzick, & Turkin (2013) argue that testing accommodations influence a student's test score in an ambiguous manner. If left out, accommodations could introduce measurement error into the VAM scores. Teachers' VAMs are estimated by the coefficients on a set of teacher fixed effects (**T**), and ε is an error term. This specification was chosen based on a detailed review of the current best practices in VAM modeling, summarized in Koedel, Mihaly, and Rockoff (2015).

We use teacher-year models instead of models that pool data over time because we are interested in teachers' effectiveness in the specific year they taught each student, not teachers' average VAM over time. As a robustness check, we also examined how our value-added estimates varied across other commonly used alternative specifications (Herrmann, Walsh, Isenberg, & Resch, 2013; Koedel et al., 2015). Our preferred model is consistently highly correlated with these alternative specifications (0.94 or above).

Since previous standardized test scores are required to calculate VAMs, we are only able to calculate VAMs for fourth through eighth grade teachers. 45% of our initial sample has a valid math VAM, or 395,426 student-year observations. Following Goldhaber et al. (2015), in our teacher quality gap models (equations 1 and 2) we use each teacher's VAM estimate from the prior school year so that students' current test scores are not taken into consideration.

Our estimates constrain the teacher fixed effect estimates to sum to zero so that teachers are compared to the average score instead of an omitted teacher. Each observation is weighted by the share of trimesters or semesters in the year during which the student-teacher link was observed.

Appendix C. LAUSD's Multiple Measures Teacher Selection Process (from Bruno and Strunk, 2019)

"Since SY2014-2015, teacher applications are processed through the following sequence. Applications are first checked for completeness and if they meet the minimum criteria. Applicants are disqualified if their application is incomplete, if their credentials are inadequate, or if there are no vacancies for particular positions. For those who pass the first round, LAUSD reaches out for profession references and ask candidates to complete an online written assessment, asking teachers to describe how they would respond to a series of different vignettes. Applicants who do not provide professional references, receive an "ineffective" rating from their references, or score lower than 11 points on their written sample are eliminated. Remaining applicants are invited to the district office for a structured formal interview and to provide sample lesson demonstrations, which are scored by HR specialists. Initial applications are scored (based on undergraduate grade point average, subject matter preparation, and background) and added to the overall applicant score from the interview, professional references, sample less, and writing sample. Applicants who receive at least 80 points and meet all the minimum required scores for each component (detailed in the Table below) are placed on an eligibility list. However, there are two possible exceptions. One, school principals can request that a specific candidate receive an exception to a score requirement (thought they still have to go through the application process). Two, candidates who fail to meet the minimum score requirement for one component, or only do not meet the 80-point requirement, are resubmitted to an HR specialist panel for a blind review. If the panel agrees that the candidate is high-quality, the candidate is added to the eligibility list. Schools draw from the eligibility list to hire for their vacancies and have flexibility in how they wish to interview/screen these candidates. School administrators never obtain the hiring scores, they only know if the candidate is eligible for hire."

Eligibility Criteria for Prospective Teachers in LAUSD

Criterion	Description	Minimum Points Possible	Maximum Points Possible	Minimum Passing Score
Interview	*Structured, conducted by one HR specialist.*	0	25	20
Professional References	*Collected from student teaching or other past professional experience.*	0	20	16
Sample Lesson	*Delivered to and evaluated by two HR specialists.*	0	15	11
Writing Sample	*Timed (45 minutes) responses to hypothetical student-related scenarios.*	1	15	11
GPA	*Scored based on verified undergraduate GPA.*	1	10	N/A
Subject Matter	*Based on subject-matter licensure test scores or, if waived, GPA score.*	8	10	N/A
Background	*For any of: certain prior LAUSD (non-teaching) experience, prior leadership (e.g., military experience), possession of a graduate degree, or Teach for America experience.*	0	2	N/A
Preparation	*For any of: attendance at school highly-ranked by U.S. News & World Report, evidence of prior teaching effectiveness (e.g., student achievement data), or major in credential subject field or, if multi-subject, core academic subject/liberal arts.*	0	3	N/A
Overall		10	100	80

Note. Points are awarded in accordance with criterion-specific rubrics aligned to district goals (e.g., employee evaluation criteria).
Applicants may be placed on the eligibility list despite scoring below the minimum passing score at the request of a school administrator or upon a review of application materials by human resources staff. (Bruno and Strunk 2019)