CEDR Policy Brief | PB #2016-9

CEDR
Center for Education Data & Research

Reconciling Different Estimates of Teacher Quality Gaps Based on Value Added

Dan Goldhaber, American Institutes for Research, University of Washington

Vanessa Quince, University of Washington

Roddy Theobald, American Institutes for Research

The Center for Education Data & Research
University of Washington Bothell
3876 Bridge Way N. Ste. 201 Seattle, WA 98103
(206) 547-5585
www.cedr.us

Abstract: This policy brief reviews evidence about the extent to which disadvantaged students are taught by teachers with lower value-added estimates of performance, and seeks to reconcile differences in findings from different studies. We demonstrate that much of the inequity in teacher value added in Washington state is due to differences across different districts, so studies that only investigate inequities within districts likely understate the overall inequity in the distribution of teacher effectiveness because they miss one of the primary sources of this inequity.

This work is supported by the William T. Grant Foundation (grant #184925) and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) (grant #R305C120008). We wish to thank James Cowan, Philip Gleason, Eric Isenberg, and Jeffrey Max for helpful comments. The views expressed in this brief do not necessarily reflect those of American Institutes for Research or the University of Washington.

Suggested citation:
Goldhaber, D., Quince, V., and Theobald, R. (2016). Reconciling Different Estimates of Teacher Quality Gaps Based on Value Added. CEDR Policy Brief 2016-9. University of Washington, Seattle, WA.

1. Evidence on Teacher Quality Gaps

A significant body of education research dating back to the Coleman Report (Coleman, 1967) documents inequity in the schooling opportunities of advantaged and disadvantaged students in U.S. public schools. More recent empirical research has documented the extent to which there are "teacher quality gaps" (TQGs) in the distribution of teacher qualifications between advantaged and disadvantaged students (both within schools and across schools and districts),[1] while an even newer line of research documents the distribution of teacher quality according to estimates from value-added models (VAMs) of teacher effectiveness.[2]

Consistent with evidence based on teacher qualifications, most of the literature focusing on value-added-based TQGs finds statistically-significant TQGs between advantaged and disadvantaged students (e.g., Chetty et al., 2014; Goldhaber et al., 2015, 2016; Isenberg et al., 2013, 2016; Mansfield, 2015; Sass et al., 2010; Steele et al., 2015). For instance, a new CALDER working paper (Goldhaber et al., 2016) uses longitudinal data from North Carolina and Washington to demonstrate that TQGs exist in both states, in every year of available data, and under different definitions of teacher quality (experience, licensure test scores, and value added) and student disadvantage (economic or minority).

However, the estimated magnitudes of the TQGs reported across this literature vary substantially from study to study. For example, a recent Mathematica final project report (Isenberg et al., 2016) uses data from 26 large school districts across the country to compare the average value added of teachers of students who are and are not eligible for free or reduced priced lunch (FRL) and concludes that "High- and low-income students have similar chances of being taught by the most effective teachers and the least effective teachers" (Isenberg et al., 2016, p. xvi). While this characterization of the magnitudes of TQGs is similar to some other studies (e.g., Chetty et al., 2014; Mansfield, 2015),[3] it conflicts with others; Steele et al. (2015), for example, argue that the TQGs reported in their study "are large enough to have meaningful consequences for children's opportunity to learn" (Steele et al., 2015, p. 94).

Some of the disagreement in the literature may simply be a matter of authors' perspectives on the importance of gap magnitudes. Mansfield (2015), for instance, finds that "students at the bottom (top) decile of a student background index are taught by teachers who are, on average, at the 41st (57th) percentile of the value-added distribution", and draws the conclusion that "teacher quality is fairly equitably distributed both within and across high schools" (Mansfield, 2015, p. 751). By contrast, Goldhaber et al. (2015) find high school TQGs of a similar magnitude and describe them as "striking" (Goldhaber et al., 2015, p. 304).

---

[1] For example, studies from states like New York (Lankford et al., 2002), North Carolina (Clotfelter et al., 2005), and Washington (Goldhaber et al., 2015) demonstrate that disadvantaged students (e.g., poor or minority) are more likely to be taught by teachers with lower qualifications (e.g., experience and licensure test scores).

[2] Value-added estimates are derived from statistical models that seek to isolate the contributions of individual teachers toward students' test achievement. For a general review of value-added methodology and its applications, see Koedel et al. (2015).

[3] Note that this characterization is not consistent with the interim report from the same project, which uses data from largely the same districts as the final project report and concludes that "On average, disadvantaged students had less access to effective teaching in the 29 study districts in grades 4 through 8" (Isenberg et al., 2013, p. xv).

Another potential explanation for seeming discrepancies in TQG findings across studies is that value added is a derived measure. Thus, unlike other measures of teacher quality such as teacher experience and licensure test scores, the value-added estimates can be sensitive to the specification of the statistical model (e.g., Goldhaber and Theobald, 2012). Indeed, Isenberg et al. (2016) argue that differences between their results and the results from much of the prior literature (including their interim project report) are primarily due to the fact that they use a model that accounts for "peer effects"—i.e., they control for the aggregated characteristics of a student's classmates (e.g., their average prior performance or the percent eligible for FRL)—and most prior research does not and thus misattributes the effect of a student's classmates to differences in teacher quality between different kinds of students.[4]

In this research brief, we explore the extent to which estimates of TQGs in Washington state are sensitive to VAM specifications, as well as a number of other possible explanations that could help reconcile different findings from the literature discussed above.

2. Potential Sources of Different Findings

To begin exploring the potential explanations for different findings about TQGs, we focus on four papers that provide directly comparable estimates of the average difference in teacher value added between non-FRL and FRL students—the Mathematica interim project report (Isenberg et al., 2013), the prior study from Washington (Goldhaber et al., 2015), the Mathematica final project report (Isenberg et al., 2016), and the new CALDER working paper using data from Washington and North Carolina (Goldhaber et al., 2016). **Table 1** summarizes both the reported gaps and the differences in the methodologies used in the papers. The estimates in column 10 of Table 1 can be interpreted as each paper's estimate for a given subject (column 8) and grade level (column 9) of the expected difference in standard deviations of student test performance between non-FRL and FRL students that is due *solely* to differences in teacher quality between non-FRL and FRL students.

The first thing to notice from Table 1 is that all studies find positive TQGs, indicating that economically-disadvantaged students are, on average, taught by lower value-added teachers. But the recent Mathematica report (Isenberg et al., 2016) reports notably smaller gaps than the other reported studies. The patterns in Table 1 are also consistent with the claim in Isenberg et al. (2016) that the smaller gaps reported in that paper may be due to the additional controls for peer effects, as this is the only one of the four studies that include these controls (see column 5).

However, three other potentially important differences between these papers could also contribute to difference in TQG findings:

[4] It remains an open question whether VAMs that account for peer effects provide less biased estimates of teacher effects. Kane et al. (2013) find that estimates from VAMs that do not account for peer effects are more predictive of future student performance (i.e., have less "forecast bias") than estimates from VAMs that do, while Chetty et al. (2014) find that a VAM with student-, classroom-, and school-level test score lags has less forecast bias than a model with just student-level lagged test scores.

1. **Setting:** Differences in average value added between non-FRL and FRL students are likely to vary depending on the school setting (column 7) considered. For example, Isenberg et al. (2016) and Goldhaber et al. (2016) both demonstrate that TQGs vary across different kinds of districts (e.g., by size and student demographics).[5]
2. **Types of gaps:** The studies from Washington and North Carolina (Goldhaber et al., 2015, 2016) rely on state-level data and therefore make comparisons both within and across districts, while the Mathematica studies (Isenberg et al., 2013, 2016) focus on individual districts and therefore make comparisons only within districts (column 2).
3. **VAM specification:** Beyond the issue of controlling for peer effects, the studies also make different decisions about whether to consider VAM estimates that include student data from the current school year (column 3) and whether to estimate VAMs that pool data across multiple years or consider each school year separately (column 4).

Fortunately, the data in Washington allow us to explore more thoroughly the potential reasons for different TQG findings. In the next section, we describe whether the difference between the findings in the four papers discussed above can be reconciled based on the type of gap measured or specification of the model used to generate value added. We cannot directly test whether the setting influences differences in findings, but we return to this issue in the conclusion.

3. Reconciling Different Findings

The student and teacher data we use to assess TQGs, and their sensitivity to different means of calculating them, come from the 2009–10 through 2012–13 school years in Washington (see Goldhaber et al., 2015, 2016 for more detail on the datasets).[6] We focus on these school years because explicit student-teacher links are available in these years of data, and as Isenberg et al. (2016) argue, models with classroom controls are most defensible when explicit student-teacher links exist. We calculate TQGs between non-FRL and FRL students based on student classroom assignments in the 2012–13 school year, though we use up to three prior years of data to estimate different VAM specifications.[7]

---

[5] Because Isenberg et al. (2016) restrict the sample from Isenberg et al. (2013) to districts and grade levels that have explicit student-teacher links, the sample of district-grade combinations in this report is smaller than the sample in Isenberg et al. (2013) and only includes elementary grades for 12 of the 26 districts.

[6] Data come from Washington state's Comprehensive Education Data and Research System (CEDARS). CEDARS data include fields designed to link students to their individual teachers, based on reported schedules. However, limitations of reporting standards and practices across the state may result in ambiguities or inaccuracies around these links. We do not use data from 2013–14 or 2014–15 because of the added complication that many schools in Washington (about one-third) participated in a pilot of the new Smarter Balanced Assessment in 2013–14, and the state did not collect scores from students in these schools (so these students are missing current-year scores in 2013–14 and prior-year scores in 2014–15).

[7] Single-year current-year VAMs are estimated from the 2012-13 school year, single-year prior-year VAMs are estimated from the 2011-12 school year, pooled current-year VAMs are estimated from the 2009-10 through 2012-13 school years, and pooled prior-year VAMs are estimated from the 2009-10 through 2011-12 school years. Following Isenberg et al. (2016), we use all four years of data to estimate the classroom coefficients in all peer effects models. For additional methodological details, see: Goldhaber et al. (2016), p. 12, Eq. 1, for the pooled-year VAM specification without peer effects; Goldhaber et al. (2015), p. 300, for the single-year VAM specification without peer effects; Isenberg et al. (2016), p. B-5, Eq. B.1, for the pooled-year VAM specification with peer

Table 2 reports the results of this exploration. We consider four different subject-grade combinations (Panels A–D) and eight different VAM specifications (for each combination of controlling for peer effects or not, using a current-year or prior-year VAM estimate, and using a single-year or pooled-year estimate).[8] Finally, to explore the influence of the types of gaps measured, we follow the procedure described in Goldhaber et al. (2015) to "decompose" the overall gap into a "District Share" (i.e., the portion of the gap due to differences in average value added across different districts), a "School Share" (i.e., the portion of the gap due to differences in average value added across different schools within the same district), and a "Classroom Share" (i.e., the portion of the gap due to differences in average value added across different classrooms within the same school).

Although Table 2 has a number of interesting patterns, we focus on two broad trends within the table. First, the "District Share" represents a substantial portion of the overall TQG in nearly every VAM specification, subject, and grade level. Thus, much of the inequity in value added in Washington state is due to differences across different school districts; i.e., districts that serve more FRL students tend to have lower average teacher quality as measured by value added.[9] This dimension of inequity is not captured by the Mathematica studies that only make comparisons within districts, and the within-district gaps in Table 2 (the sum of "School Share" and "District Share") are in fact relatively comparable to within-district gaps reported in Isenberg et al. (2016).

Second, the influence of classroom controls on estimated TQGs in terms of value added is much stronger in middle school grades (Panels C and D) than in elementary grades (Panels A and B). Specifically, while the estimated gaps in elementary grades become marginally larger when the VAM controls for aggregated classroom characteristics, the estimated gaps in middle school grades decrease substantially when the VAM includes classroom controls. A potential explanation for this difference is the prevalence of student tracking in higher grade levels (e.g., Jackson, 2014), which may make aggregated classroom characteristics more predictive of student performance in middle school than in elementary school.

4. Conclusions and Open Questions

This preliminary analysis suggests that the different findings from the prior literature summarized in Table 1 may be reconciled by the following observations:

---

effects; and Isenberg et al. (2016), pp. B-5–B-6, Eqs. B.1–B.3, for the single-year VAM specification with peer effects.

[8] We also estimated variants of these VAMs that controlled for teacher experience, and found that these gaps tend to be smaller but still positive, meaning that TQGs are not solely due to the disproportionate assignment of FRL students to novice teachers. See Goldhaber et al. (2016) for additional details.

[9] As reported in Goldhaber et al. (2016), we also investigate TQGs in the 34 districts in the Puget Sound Education Service District, since these districts are substantially larger than the average district in the state and are more similar demographically to the districts considered in Isenberg et al. (2016). We find that cross-district TQGs in these districts are an even larger percentage of the overall TQG in these districts than across the state as a whole, and that the within-district TQGs in these districts are comparable to those reported in Isenberg et al. (2016).

- Much of the inequity in value added (at least in Washington state) is due to differences across different districts, so Isenberg et al. (2016) likely understate the overall inequity in the distribution of teacher effectiveness because they focus solely on within-district gaps and thus miss one of the primary sources of this inequity (i.e., the cross-district sorting of students and teachers within a state).
- Estimating models with peer effects appears to have a substantial impact on TQGs in middle school, which likely explains some of the difference between the large middle school TQGs reported in Goldhaber et al. (2015) and the much smaller TQGs reported in Isenberg et al. (2016).[10] It is an open question whether VAMs that include peer effects produce more accurate estimates of teacher quality, so these different findings are subject to different interpretations.[11]

The primary open question that stems from this exploration is whether these patterns hold in different settings. In fact, given the evidence in Goldhaber et al. (2016) and Isenberg et al. (2016) that there is considerable heterogeneity in TQGs across different districts, we would *expect* to find different answers in different settings. We therefore urge further research into teacher quality gaps in U.S. public schools that considers the implications of different measures of teacher quality (including value added) on the estimation of these gaps.

[10] This is particularly plausible because the Isenberg et al. (2016) sample disproportionately consists of students in grades 6–8.

[11] VAMs with peer effects are conceptually appealing, but there are many different specifications of peer-effect VAMs (e.g., Chetty et al., 2014; Isenberg et al., 2016; Kane et al., 2013) and each has different strengths and weaknesses (Koedel et al., 2015). For example, the specification in Isenberg et al. (2016), which does not control for teacher experience in estimating peer effects, may lead to biased estimates of TQGs if teachers tend to teach in more advantaged classrooms as they gain early-career teaching experience (in which case returns to teaching experience could be misattributed to peer effects).

References

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392.

Coleman, J. S. (1967). *The concept of equality of educational opportunity.* Washington, DC: U.S. Office of Education.

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. (2016). *Has it always been this way? Tracing the evolution of teacher quality gaps in U.S. public schools.* CALDER Working Paper 171.

Goldhaber, D., & Theobald, R. (2012). *Do different value-added models tell us the same things?* Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/

Isenberg, E., Max, J., Gleason, P., Johnson, M., Deutsch, J., & Hansen, M. (2016). *Do low-income students have equal access to effective teachers? Evidence from 26 districts.* NCEE 2017-4008. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/pubs/20174008/pdf/20174008.pdf

Isenberg, E., Max, J., Gleason, P., Potamites, L., Santillano, R., Hock, H., & Hansen, M. (2013). *Access to effective teaching for disadvantaged students.* NCEE 2014-4001. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/pubs/20144001/pdf/20144001.pdf

Jackson, C. K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645–684.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://eric.ed.gov/?id=ED540959

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

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

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, 33(3 Part 1), 751-788.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2010). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2), 104–122.

Steele, J. L., Pepper, M. J., Springer, M. G., & Lockwood, J. R. (2015). The distribution and mobility of effective teachers: Evidence from a large, urban school district. *Economics of Education Review*, 48, 86-101.

Table 1. Average Difference in Teacher Value Added Between Non-FRL and FRL Students in Published Literature

1	2	3	4	5	6	7	8	9	10
Paper	Type of Gaps	VAM Year	VAM Spec	VAM Peer Effects	Years	Setting	Subject	Grades	Gap
Isenberg et al. (2013)	Within districts	Current year	Single year	No	2008–09 to 2010–11	29 School districts	ELA	4–8	0.034
							Math	4–8	0.024
							ELA	4–5	0.040
							Math	4–5	0.021
							ELA	6–8	0.030
							Math	6–8	0.027
Goldhaber et al. (2015)	Within and across districts	Prior year	Single year	No	2011–12	Washington state	Combined	4	0.035
							ELA	7	0.037
							Math	7	0.059
Isenberg et al. (2016)	Within districts	Current year	Single year	Yes	2008–09 to 2012–13	26 Districts (12 with grades 4–8; 14 with grades 6–8)	ELA	4–8	0.005
							Math	4–8	0.005
							ELA	6–8	0.012
							Math	6–8	0.012
Goldhaber et al. (2016)	Within and across districts	Current and prior years	Pooled across current and prior years	No	2006–07 to 2012–13	Washington state	Combined	4–5	0.025 to 0.033
					1998–99 to 2012–13	North Carolina	Combined	4–5	0.024 to 0.031

Notes: ELA, English language arts; FRL, free or reduced-price lunch; Spec, specification, VAM, value-added model

Table 2. Average Difference in Teacher Value-Added Between FRL and Non-FRL Students in Washington, 2012–13

Panel A: Grades 4–5 Math								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Total gap	0.023	0.024	0.018	0.023	0.024	0.028	0.026	0.038
District share	0.018	0.018	0.014	0.015	0.017	0.018	0.017	0.021
School share	0.002	0.002	0.001	0.003	0.004	0.005	0.006	0.010
Classroom share	0.002	0.003	0.003	0.005	0.003	0.005	0.004	0.007
Prior year (vs. current year)			X	X			X	X
Peer effects (vs. no peer effects)		X		X		X		X
Pooled spec (vs. single year)					X	X	X	X
Panel B: Grades 4–5 ELA								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Total gap	0.032	0.046	0.020	0.033	0.030	0.044	0.024	0.043
District share	0.022	0.029	0.014	0.021	0.021	0.028	0.017	0.026
School share	0.009	0.013	0.005	0.010	0.008	0.013	0.006	0.013
Classroom share	0.002	0.004	0.001	0.002	0.002	0.003	0.002	0.003
Prior year (vs. current year)			X	X			X	X
Peer effects (vs. no peer effects)		X		X		X		X
Pooled spec (vs. single year)					X	X	X	X
Panel C: Grades 6–8 Math								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Total gap	0.065	0.032	0.044	0.015	0.056	0.027	0.047	0.032
District share	0.029	0.012	0.012	-0.003	0.026	0.010	0.020	0.012
School share	0.010	0.003	0.008	0.001	0.007	0.001	0.007	0.004
Classroom share	0.026	0.017	0.023	0.016	0.023	0.016	0.020	0.016
Prior year (vs. current year)			X	X			X	X
Peer effects (vs. no peer effects)		X		X		X		X
Pooled spec (vs. single year)					X	X	X	X
Panel D: Grades 6–8 ELA								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Total gap	0.035	0.007	0.024	-0.001	0.037	0.010	0.032	0.017
District share	0.016	0.000	0.009	-0.006	0.016	0.000	0.013	0.004
School share	0.007	0.000	0.006	0.001	0.008	0.002	0.008	0.004
Classroom share	0.012	0.007	0.008	0.005	0.013	0.009	0.011	0.009
Prior year (vs. current year)			X	X			X	X
Peer effects (vs. no peer effects)		X		X		X		X
Pooled spec (vs. single year)					X	X	X	X

Notes: ELA, English language arts; FRL, free or reduced-price lunch.