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

The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce

CORY KOEDEL AND
JIAXI LI

WORKING PAPER 106 • JANUARY 2014

The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce

Cory Koedel
University of Missouri - Columbia

Jiaxi Li
University of Missouri - Columbia

Contents

Acknowledgements

This research was supported by the National Center for Analysis of Longitudinal Data in Education Research (CALDER) funded through Grant R305C120008 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education.

 The authors thank Eric Isenberg and Eric Parsons for useful comments.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce

Cory Koedel and Jiaxi Li

CALDER Working Paper No. 106
January 2014

Abstract

We examine the efficiency implications of imposing proportionality in teacher evaluation systems. Proportional evaluations force comparisons to be between equally-circumstanced teachers. We contrast proportional evaluations with global evaluations, which compare teachers to each other regardless of teaching circumstance. We consider a policy where administrators use the ratings from the evaluation system to help shape the teaching workforce, and define efficiency in terms of student achievement. Our analysis indicates that proportionality can be imposed in teacher evaluation systems without efficiency costs under a wide range of evaluation and estimation conditions. Proportionality is efficiency-enhancing in some cases. These findings are notable given that proportional teacher evaluations offer a number of other policy benefits.

1. Introduction

State and local education agencies across the United States are working to improve teacher quality through the adoption of rigorous teacher evaluation systems.[1] The teacher-performance signals that come out of these systems can be acted on in a number of ways to improve outcomes for students in K-12 schools. However, despite the rapid growth in the development of teacher evaluation systems nationwide, there is still much controversy surrounding the specifics of how to measure teacher performance. The lack of consensus in this area is reflected in the variety of different approaches that state and local education agencies use to evaluate teachers.

This paper contributes to the literature by examining the efficiency effects of using different evaluation metrics to rank-order teachers with the objective of using the rankings to help shape the teaching workforce. We perform our analysis using simulated data and measure efficiency in terms of student achievement. The simulated data are constructed following the literature on test-based measures of teacher performance because the properties of test-based measures are well understood, at least relative to available alternatives (e.g., classroom observations, student evaluations).[2] However, the substance of our findings will apply to any direct measure of teacher performance, including those commonly used in the "combined measures" that are being developed by a number of state and local education agencies (e.g., Bill and Melinda Gates Foundation, 2013; Mihaly et al., 2013; Strunk, Weinstein and Makkonnen, 2013).

[1] A number of states have already enacted legislation mandating performance-based evaluations, and high stakes have been attached in some cases. For example, Senate Bill 736 in Florida (2011) and House Bill 1001 in Colorado (2012) are examples of legislation linking high stakes decisions to performance-based teacher evaluations. Similar legislation is being considered or is in the process of being implemented in other states, including Michigan and Pennsylvania. Some large school districts are also independently developing performance-based teacher evaluation systems. The Houston Independent School District (Shifrer, Turley and Heard, 2013), Los Angeles Unified School District (Strunk, Weinstein and Makkonnen, 2013), Pittsburgh Public School District (Chute, 2013), and Washington DC Public School District (Arcaira et al., 2013) are examples.

[2] There is a vast literature examining test-based performance measures and their properties – examples of studies include Ehlert et al. (2013a, 2013b), Goldhaber, Walch and Gabele (2013); Goldhaber and Hansen (forthcoming); Koedel and Betts (2007); McCaffrey et al. (2009); and Sass, Semykina and Harris (forthcoming). Hanushek and Rivkin (2010) provide a recent overview of the test-based literature; also see McCaffrey et al. (2003). Researchers are just beginning to rigorously explore the properties of non-test based measures (e.g., see Polikoff, 2013).

We build on the basic simulation framework provided by Winters and Cowen (2013) to generate the data for our simulations. We compare evaluation systems that identify and rank order teachers based on (1) proportional estimates of teacher quality, which force comparisons to be between equally-circumstanced teachers (Ehlert et al., 2013a), and (2) global estimates of quality that compare teachers to each other regardless of teaching circumstance.[3] We compare the systems within the context of a policy that uses the teacher rankings to help shape the workforce. In particular, we consider a removal policy targeted at the bottom 10 percent of teachers.

Because of the conditional nature of the proportional estimates, teacher rankings based on these estimates need not be consistent with rankings based on global estimates of quality. To illustrate how the rankings can differ, consider the following example: suppose that there are two types of schools, type-A and type-B, and that teacher quality is higher in type-A schools.[4] A quality-based removal policy that depends on global rankings will identify more teachers in type-B schools to be removed. In contrast, an analogous policy based on teachers' proportional rankings will ensure that an equal number of teachers from type-A and type-B schools are removed.

We examine the efficiency effects of the different policies only in terms of how they influence which teachers are removed and replaced. We do not allow the proportional policy to otherwise improve workforce quality – for example, we do not allow proportionality to increase educator effort,

[3] Ehlert et al. (2013a) identify three key objectives of evaluation systems in education and argue that proportional rankings are the most desirable given these objectives. Discussions of the proportionality principle – although it is referred to by different names in different contexts – can also be found in the economics literature (e.g., see Barlevy and Neal, 2012; Schotter and Weigelt, 1992).

[4] Type-A schools can be thought of as low-poverty schools and type-B schools can be thought of as high-poverty schools. This example is motivated by empirical evidence showing gaps in teacher quality between high- and low-poverty schools (e.g., see Arcaira et al., 2013; Goldhaber, Walch and Gabele, 2013; Isenberg et al., 2013; Sass et al., 2012). There are also a number of studies that discuss the general recruiting challenges faced by high-poverty schools (Boyd et al., 2005; Clotfelter et al., 2006; Jacob, 2007; Reininger, 2012).

nor do we allow for the possibility that proportional performance signals foster more productive educator learning (Ehlert et al., 2013a).[5] To the extent that these aspects of proportional evaluations also improve efficiency, our findings likely understate the efficiency gains from proportional policies.

Focusing strictly on the workforce-shaping effects of the policies, it is straightforward to show that it is more efficient to use a proportional policy when there is a gap in average quality between teachers who teach in different schooling contexts, which recent research suggests is likely (e.g., see Arcaira et al., 2013; Goldhaber, Walch and Gabele, 2013; Isenberg et al., 2013; Sass et al., 2012). The key insight underlying the efficiency gain from proportionality is that the effect of a targeted removal policy will depend not only on the quality of the teachers being removed, but also on the quality of replacement teachers. Continuing with the example from above, note that under plausible conditions the gap in quality between teachers in type-A and type-B schools will persist for potential replacement teachers at these schools as well. After taking direct account of the link between observed teacher quality and the quality of teacher replacements for schools in different contexts, we show that the proportional policy is the most efficient in terms of raising student achievement.

Although the efficiency rationale for proportionality is compelling, our analysis indicates that in real-world applications the likely benefits from imposing proportionality will be small. One reason is that evaluating teachers in practice requires the use of imprecise measures, which attenuates the efficiency effect of the proportional policy. Another is that the efficiency gains from proportionality can be offset by gaps in the *variance* of teacher quality across different types of schools. However, it is important to recognize that as long as the proportional policy performs no worse than available alternatives, it may be preferred by policymakers because it offers other benefits. Ehlert et al. (2013a) provide a detailed discussion of several benefits of proportionality, which we briefly review below. We also discuss an

[5] We do not directly account for these potential benefits of the proportional policy because there is not a sufficient research literature to draw on to quantify, and thus parameterize in our simulations, the benefits of proportionality along these dimensions.

additional pragmatic benefit that is likely to become increasingly important as performance-based teacher evaluations come online at scale: proportional policies can be used to assuage concerns from labor groups about fairness (Polikoff et al., forthcoming; also see Vaznis, 2013).

The remainder of this paper is organized as follows. Section 2 provides general background information and policy context. Section 3 describes the construction of the simulated data. Section 4 illustrates the efficiency gains from the proportional policy under straightforward evaluation conditions. Section 5 examines the robustness of the efficiency result to a variety of complications to the data generating process and estimation procedure, and examines the issue of equity. Section 6 discusses mechanisms for the link between observed quality in K-12 schools and the quality of replacement teachers – this link is central to our findings. Section 7 extends the intuition from our analysis to non-test-based performance measures for teachers. Section 8 addresses other considerations that are important to policymakers and generally favor the use of proportional evaluations. Section 9 concludes.

2. Background

Motivation for Improving Teacher Evaluation Systems

A large research literature shows that teachers differ dramatically in their effectiveness as measured by value-added to student test scores (for a recent overview see Hanushek and Rivkin, 2010). Furthermore, Chetty, Friedman and Rockoff (2011) link differences in exposure to effective teachers, as measured by value-added, to differences in later-life outcomes for students. The consistency of the empirical evidence regarding the importance of teacher quality, combined with the difficulty that researchers have had linking performance differences between teachers to observable characteristics (Kane, Rockoff and Staiger, 2008; Nye, Konstantopoulos and Hedges, 2004; Rivkin, Hanushek and Kain, 2005), motivates the incorporation of direct, outcome-based performance measures into teacher

evaluations. Recent evidence from Dee and Wyckoff (2013) suggests that workforce quality can be improved through the careful implementation of educator evaluation systems.[6]

As noted above, a number of state and local education agencies have intensified efforts around the construction and use of performance-based measures for teacher evaluations. Table 1 in Winters and Cowen (2013) provides a recent overview at the state level. Most agencies are constructing what have come to be called "combined measures" of teacher performance. Combined measures typically include achievement-based performance metrics, classroom observations, student surveys, etc. (Bill and Melinda Gates Foundation, 2013; Dee and Wyckoff, 2013; Mihaly et al., 2013; Strunk, Weinstein and Makkonnen, 2013).[7]

Although performance-based teacher evaluations are increasingly common, and increasingly associated with high-stakes decisions, the research literature is thin in terms of specific guidance for constructing the evaluation metrics. A number of studies exist that compare output from alternative models, and discuss potential tradeoffs, but these studies typically do not offer concrete guidance pointing to a clear course of action (e.g., see Goldhaber, Walch and Gabele, 2013; Goldhaber, Goldschmidt and Tseng, 2013; Ehlert et al., 2013b). Along some dimensions it is not reasonable to expect a clear course of action to emerge – many of the modeling choices that must be made by administrators come with tradeoffs that require decisions along inherently subjective dimensions. However, along other dimensions the research literature can be more definitive. Our study aims to inform the decision-making process by providing concrete evidence about a particular aspect of the model-selection process. Specifically, we examine the efficiency effects of imposing the proportionality

[6] Dee and Wyckoff (2013) evaluate the Washington, DC IMPACT program and provide regression-discontinuity evidence showing that the program improves workforce quality in several ways. The IMPACT program uses "combined measures" of teacher performance (see next paragraph).

[7] Legislation in a number of states mandates that a minimum percentage of teachers' overall ratings depend on student achievement growth. Examples of states with such mandates include Florida and Colorado.

property in teacher evaluation systems. We define efficiency in terms of student achievement – the most efficient system is the one that results in the largest improvements in achievement.

What is Proportionality?

Proportional evaluations force comparisons to be between equally-circumstanced teachers. The term "proportionality" refers to the representation of teachers throughout the rankings that emerge from the evaluation system. A strictly proportional ranking system is such that if *x* percent of the teaching population teaches in schooling environment *y* (e.g., in high-poverty schools), then *x* percent of any subset of teacher rankings (e.g., the top quintile) includes teachers who teach in schooling environment *y*.

Proportional rankings can be constructed in a number of straightforward ways. Ehlert et al. (2013a) estimate a proportional model to measure school value-added in Missouri. They use a two-step fixed effects procedure. The key functional feature of their approach is that it partials out the variance in student test scores attributable to the observable characteristics of students and schools prior to estimating the value-added measures. Also note that proportionality can be enforced outside of the model of student achievement. For example, one can perform *ex post* regression adjustments at the unit of evaluation (e.g., district, school, teacher).[8] We refer the interested reader to Ehlert et al. (2013a) for more information.

Enforcing the proportionality property in teacher rankings can have meaningful evaluative consequences in cases where teacher quality differs systematically across different types of schools. For example, the previous literature shows that estimated teacher quality is consistently lower in high-poverty schools. Recent studies that illustrate this empirical regularity include Goldhaber, Walch and

[8] In fact, strict proportionality of teacher rankings need not hold if the observable student and school characteristics are controlled for only in the student-achievement specification – this is because of weighting issues. See Ehlert et al. (2013a) for details.

Gabele (2013), Isenberg et al. (2013), and Sass et al. (2012); also see Clotfelter et al. (2006). We take the previously-documented gap in estimated teacher performance between high- and low-poverty schools as given and consider cases where (1) the gap reflects a causal difference in teacher performance across school types and (2) the gap also partly reflects bias generated by the inability of available models to adequately control for teaching circumstance. We primarily compare proportional rankings to global rankings under the former condition. This is the best-case scenario for the use of global rankings.[9]

3. Generating the Data

Our simulated-data framework is constructed based on previous work by Winters and Cowen (2013). We specify each teacher's annual estimated performance measure as the sum of three components:

$$\hat{\gamma}_{jt} = q_j + \delta_{jt} + \eta_{jt} \qquad (1)$$

In equation (1) q_j is a time-invariant performance measure, δ_{jt} varies from year to year and is independent across years (δ_{jt} can be viewed as reflecting year-to-year teacher-classroom match effects and/or natural variation in teacher performance from one year to the next), and η_{jt} is a residual component that reflects sampling variance attributable to the draw of students for teacher *j* in year *t*. One small way that our setup deviates from that of Winters and Cowen is that they draw their analog to η_{jt} without actually assigning students to teachers – this is because there are no students in their simulations. In our setup we generate student-level data so that η_{jt} truly reflects sampling variance.

[9] If the estimated teacher-quality gap between high and low-poverty schools is driven by inadequate controls, then the proportional model would likely be preferred for other reasons beyond those discussed in this paper. For example, consider the scenario where there is no real quality gap and estimated gaps entirely reflect bias. In this scenario, proportional evaluations would be preferred for their benefits in terms of bias reduction in addition to the reasons discussed in this paper.

We generate student test scores as follows after randomly assigning students to teachers:[10]

$$Y_{ijt} = \alpha_{it} + (q_j + \delta_{jt}) \quad (2)$$

Equation (2) shows that student scores are a function of a student-year specific component, α_{it}, and teacher assignments. α_{it} captures a number of factors that influence student test scores, perhaps most notably student ability and test measurement error (Boyd et al., 2012; Koedel, Leatherman and Parsons, 2012). For the purposes of our application, α_{it} is best viewed as the residual variance in student test scores after accounting for the role of teachers. Note that equations (1) and (2) are linked through α_{it} because $\eta_{jt} = \frac{1}{N_{jt}} \sum_{i=1}^{N_{jt}} \alpha_{it}$, where N_{jt} is the number of students in teacher *j*'s classroom in year *t*.

Winters and Cowen use a number of different parameterizations for the components of equation (1) – for our analysis we use one of the parameterizations from their paper where $\sigma_q = 0.15$, $\sigma_\delta = 0.15$, $\sigma_\eta = 0.21$. Implicit in the Winters and Cowen setup is that the variance in student scores is normalized to one; we use the same normalization for our simulations (thus $\sigma_Y = 1.0$). With random assignment of students to teachers the expected variance of η is entirely driven by the student/teacher ratio. The above-specified variance of η is achieved when we set this ratio to 20, which along with the above-specified variances of q_j and δ_{jt} results in a year-to-year correlation in $\hat{\gamma}_{jt}$ of approximately 0.25. This parameterization uses a plausible value of σ_q and also produces plausible estimates of σ_γ

[10] We assign students to teachers randomly to maintain focus on our research question with limited distractions, as in Schochet and Chiang (forthcoming). Winters and Cowen (2013) also effectively assume random assignment of students to teachers by ensuring unbiasedness in estimated teacher quality. In cases where assignment is not random, the efficacy of the evaluation will depend in part on how well available control variables can account for the non-random assignment. Recent evidence from Chetty, Friedman and Rockoff (2011), Goldhaber and Chaplin (2012) and Koedel and Betts (2011) offers reason for optimism about available models, but a detailed discussion of this issue is beyond the scope of the present study.

(e.g., see Hanushek and Rivkin, 2010). Our findings are not qualitatively sensitive to reasonable alternative parameterizations.

Thus far we have laid the foundation for our simulations based on previous work. We take this foundation as a point of departure and add differentiated schools.[11] We group teachers into one of two school types: (1) type-A (low-poverty schools) and (2) type-B (high-poverty schools). We introduce the school types so that we can incorporate the empirical regularity that in the absence of forced proportional comparisons, estimated teacher quality for teachers in low-poverty schools, on average, is higher than estimated teacher quality for teachers in high-poverty schools.[12]

We generate data for 12,000 students and 600 teachers, which results in a student-teacher ratio of 20. We do not allow the student-teacher ratio to vary across teachers. Teachers are divided into two groups of 300 where the first group teaches in type-A schools and the second group in type-B schools. We do not allow school effects to enter into the data generating process directly in any way. For example, there are no principal effects, and the student component of test scores is drawn for all students in all schools from the same distribution, $\alpha \sim N(0, \sigma_\alpha)$. Although our simulation framework was initially designed to allow for heterogeneous school effects and student sorting, later it will become clear that formal analysis is not required to extend our findings to cases where schools are heterogeneous and/or students are sorted across schools.

4. Basic Results

We begin by drawing teachers in type-A and type-B schools from the quality distributions $q_j^A \sim N(0.05, 0.15)$ and $q_j^B \sim N(-0.05, 0.15)$, respectively. Thus, the expected gap in quality between

[11] Winters and Cowen (2013) briefly consider a scenario with differentiated schools; however, they do not consider the proportionality issue in their study.

[12] We can also set up the data so that schools differ by a continuous poverty measure, such as the share of students eligible for free/reduced-price lunch or the share disadvantaged minority. We use type-A and type-B schools for ease of presentation and without loss of generality.

low- and high-poverty schools is 0.10 standard deviations of the student achievement distribution, parameterized in the first moment (we consider scenarios with more complex quality gaps in Section 5). We examine the efficiency effects of teacher removal policies where replacement teachers for vacancies in type-A and type-B schools are drawn from these same distributions. The initial distributions imply what is perhaps an implausibly large difference in quality across school types, but they are useful for illustration. We consider more moderate quality gaps later on.[13]

The Efficiency Rationale for the Proportional Policy

Table 1 shows what happens when we remove the bottom 10 percent of the teaching workforce for a single year based on global and proportional teacher rankings. Removals are based on teachers' actual values of *q*. Initially we do not allow for teacher attrition except for attrition that occurs as a direct consequence of the removal policy. The table shows the average achievement effect of each policy system-wide and the change in teacher quality in the slots where replacements occur.

Workforce quality is higher overall when teachers are removed using the proportional policy – averaged across the entire system, the achievement gain per student is 0.0250 standard deviations of student test scores under the global policy and 0.0263 under the proportional policy. The achievement gains are concentrated among students who are taught by the replacement teachers, which means that the average gain in *q* per removed teacher under each policy is equal to ten times the system-wide achievement gain: 0.250 and 0.263 for the global and proportional policies, respectively.[14]

---

[13] Imposing the quality gap across school types also increases σ_q as measured across all schools. In principle we could reduce σ_q within school type to offset the gap but doing so has no bearing on the substance of our findings and comes at the expense of tractability (particularly later on when we make adjustments to the data generating process).

[14] A symmetric result holds for policies aimed at improving retention rates for the most effective teachers (see discussion in Section 5.7). Conceptually, the retention-targeted analog to the removal policy would involve selective retention bonuses that would lower the probability of attrition for the most effective teachers. An added benefit from

Figure 1 illustrates the intuition behind the efficiency gain from proportionality, maintaining the case in Table 1 where removals are based on known *q*. Across 1,000 simulations, the figure shows the average percentile ranking and *q*-value for the last removed teacher in each school-type using the proportional and global policies. The global policy is structured so that the average difference in quality (*q*) between the marginally removed teachers in type-A and type-B schools is essentially zero. Given the overall gap in quality across school types, this result is achieved by removing a much larger share of teachers in type-B schools. Specifically, on average, the global policy removes 15.4 percent of teachers in type-B schools and 4.6 percent of teachers in type-A schools.

To illustrate the inefficiency of the global policy, consider taking the last teacher who is removed from a type-B school based on the global rankings (i.e., among the teachers who are removed in type-B schools, the teacher with the highest value of *q*), rehiring that teacher, and then removing the next teacher in line from a type-A school instead. The marginally rehired teacher in type-B schools has essentially the same *q*-value as the teacher who is removed in her place from a type-A school (as shown in Figure 1). However, the expected value of *q* for the type-A replacement teacher is much larger than that for the type-B replacement teacher. The expected net gain in *q* from making this change at the margin can be written as:

$$Gain = [E(q_A^{0.50}) - E(q_A^{0.046})] + [E(q_B^{0.154}) - E(q_B^{0.50})] \quad (3)$$

The subscripts on the terms in equation (3) indicate the school type and the superscripts indicate the percentile of the type-specific distribution of *q*. The first term in square parentheses is positive and represents the expected gain in *q* from removing the next-in-line type-A teacher. The second term in square parentheses is negative and represents the expected loss associated with rehiring the marginally-

using proportional measures to determine retention bonuses is that the efficiency gains are focused in high-poverty schools.

removed teacher in type-B schools (the loss occurs because the rehired teacher is worse in expectation than her replacement from the type-B distribution). As shown in Figure 1, $E(q_A^{0.046}) \approx E(q_B^{0.154})$, so the expected net gain based on our parameterization thus far is 0.10.[15] That is, the marginal move toward proportionality raises overall teacher quality, and thus student achievement. Subsequent marginal moves toward proportionality continue to raise student achievement up to the point where proportionality is achieved. This is the logic underlying the efficiency gain from the proportional policy.[16]

Policies Based on Unbiased Estimated q

Now we move to the case where *q* is not known but must be estimated. We assume that *q* can be estimated without bias – that is, quality rankings can be produced that deviate from the true quality rankings (based on actual *q*) only because of statistical imprecision. We can construct such a scenario with our simulated data because we randomly assign students to teachers. In real-world applications, this scenario is informative if we believe that the model from which we estimate $\hat{\gamma}$ is sufficiently rich so that there is no bias (or negligible bias) in the teacher-quality estimates.

Table 2 replicates Table 1 except that we remove teachers based on $\hat{\gamma}$ rather than q, with removals depending on single-year estimates of $\hat{\gamma}_{jt}$ for each teacher. As anticipated, the noise in the performance signal attenuates the policy impact. The average system-wide gains in achievement under the global and proportional policies in Table 2 are 0.0127 and 0.0129, respectively.[17] The efficiency gain from using the proportional policy remains despite being muted considerably by the statistical imprecision of the estimates upon which the removal decisions are based. Returning to the logic from

[15] That is, the expected gain is equal to the expected gap in quality across school-types.
[16] Although our study is primarily focused on the efficiency implications of proportionality, equity considerations also merit attention. Comparing the policies in terms of equity requires some nuance. We discuss the issue of equity in detail in Section 5.7.
[17] The general magnitude of our reported policy effects is smaller than in Winters and Cowen (2013) because they iterate their removal policy for a number of years. We consider iterative policies later on and, consistent with their work, obtain much larger policy effects.

Figure 1, the lower bound on the efficiency gain from using the proportional policy in the presence of statistical imprecision is zero. This will occur when $\hat{\gamma}_{jt}$ entirely reflects noise such that the firings are effectively random.

5. Robustness and Extensions

Allowing for Biased Estimates of Teacher Performance

In Section 4.2 we introduced noise into the estimation process, but not bias. Enforcing unbiasedness puts the global measures in the best possible light because the likely sources of bias in standard models are such that the bias will favor teachers in low-poverty (type-A) schools (e.g., see Ehlert et al., 2013a). In this section we consider the case where the estimated teacher effects are biased.

Bias will factor into the comparison between global and proportional rankings if it occurs systematically across school types – that is, if the bias favors teachers in one school-type over teachers in the other. Alternatively, while within-sector bias is a generally important concern, as long as it is consistent in direction and magnitude within sectors it will not have any bearing on whether global or proportional rankings are preferred – both will be equally affected by within-sector bias.[18]

With this in mind, we consider the case where the global rankings are biased by the inability of the statistical model to appropriately control for schooling context. The importance of controlling for context in models that estimate teacher (and school) effectiveness has been discussed in detail in previous studies including Ehlert et al. (2013a) and Raudenbush and Willms (1995). Ehlert et al. make the argument that the most likely sources of bias in models that aim to generate global rankings will

[18] Of course, if bias is an important concern than using the effectiveness estimates for high-stakes decisions may be generally undesirable. Recent evidence offers some optimism for the value of test-based measures (Chetty, Friedman and Rockoff, 2011), although non-test-based measures have been less-rigorously investigated. A discussion of how much bias would be too much is beyond the scope of this paper. Obviously, our comparison between the global and proportional policies is moot if the underlying performance measures are deemed too unreliable to be useful. Note that any bias driven by unobserved selection (e.g., better teachers being assigned to better students along unobserved dimensions, perhaps within schools) is unaddressed by either modeling approach.

favor advantaged schools and teachers teaching in these schools. The reason is that there may not be sufficient variation to properly identify the coefficients that control for schooling context, leading to attenuation in the control-variable coefficients and correspondingly, bias in the estimated teacher effects. Of course, this presumes that the evaluation system is using a model that attempts to control for context but this is not always the case. A number of states are evaluating teachers using performance measures estimated from "sparse" growth models that do not control for student or school characteristics at all (beyond prior student test scores), with "Student Growth Percentiles" (SGPs) being a particularly popular variant of this approach (Betebenner, 2009).[19]

We introduce bias into the estimated teacher effects by imposing *ad hoc* bias terms of +0.02 and -0.02 on the estimates for teachers in type-A and type-B schools, respectively. Table 3 shows results analogous to Table 2 but with the bias built into the estimates. While the real gap in teacher quality remains as before at 0.10, the estimated gap is now 0.14.

The bias terms offer additional protection for teachers in type-A schools when removals depend on the global policy. However, the influence of the bias is mitigated by the proportional policy. Unsurprisingly, the proportional model is even more efficient relative to the global-but-biased alternative. The system-wide gain in student achievement in Table 3 when removals depend on the global policy is 0.0125. Under the proportional policy the gain is 0.0130.

The general takeaway from Table 3 is as follows: when a model that aims to globally rank teachers is subject to bias, the proportional policy offers protection against inefficiency created by the bias (to the extent that the bias aligns with observable differences in schooling context). Although one

[19] The SGP literature does not support the use of these measures to estimate teacher effectiveness (Betebenner, 2009). Still, a number of states, including Colorado and Massachusetts, appear to be using them for precisely this purpose. Also note that there is nothing inherent in the SGP approach that prevents it from taking account of student and school characteristics. In fact, properly adjusted SGPs could be used in a proportional evaluation. However, as a practical matter this is not currently how SGPs are used.

might presume that bias that aligns with observable differences in schooling context can be readily removed by standard modeling approaches without enforcing proportionality, this is only true under a set of assumptions that need not be met. Ehlert et al. (2013a) provide a detailed critique of these assumptions.

Allowing for Natural Attrition

Thus far we have not allowed for teacher attrition other than attrition driven by the removal policy. In reality teacher attrition in the absence of such policies is high. In this section we layer the removal policy on top of natural teacher attrition. We consider natural attrition that depends on teacher quality as parameterized by Winters and Cowen (2013), who use estimates from Feng and Sass (2011). Feng and Sass (2011) show that the relationship between teacher attrition and quality is U-shaped so that the most and least effective teachers are the most likely to leave.

Table 4 shows how incorporating natural teacher attrition affects our results. Unlike in the previous tables, the average gain in q per removed teacher is no longer equal to ten times the average gain in achievement for all students. This is because less than 10 percent of teachers are involuntarily removed under the 10-percent removal policy. Put differently, some of the teachers targeted for removal based on the policy elect to leave anyway, which means that their exits are not part of the policy effect.

We use estimated teacher effectiveness to make the removal decisions in Table 4 (without bias), which means that the results in Table 2 serve as the baseline comparison case without attrition. Relative to Table 2, the total achievement effect of both policies is smaller because fewer teachers are removed involuntarily. There is also a small decline in the average gain in q per removal from what we show in

Table 2 using either removal policy. The most efficient policy in Table 4 continues to be the proportional policy, albeit only marginally.[20]

Using Multiple Years of Data to Inform Policy Action

We have used single-year estimates of teacher performance to determine removal decisions thus far. In practice, most evaluation systems incorporate multiple years of data to improve precision and avoid unduly penalizing or rewarding teachers for one particularly good or bad year. In this section we incorporate this dimension of real-world teacher evaluations by using estimates of teacher effectiveness based on three years of data to determine removals. We implement the three-year removal policies with and without natural teacher attrition in an otherwise static framework.

We first build the analog to Table 2 without natural teacher attrition. To do this, we hold the workforce static for three years, remove the bottom 10 percent of teachers one time after the third year, and then estimate the improvements in workforce quality at the beginning of year-4. Table 5 shows our results, which can be compared with our results in Table 2 to illustrate the value of using additional years of data to inform removal decisions. Unsurprisingly, the performance of both policies is improved using the three-year estimates. This is because the roles of δ_{jt} and η_{jt} in determining the removal decisions are reduced. The efficiency gain from the proportional policy, above and beyond the global policy, is equal to approximately to 0.0003 standard deviations of student test scores system-wide (compared to 0.0002 standard deviations of student test scores in Table 2).

To build the analog to Table 4 we allow for natural teacher attrition to occur during the first three years, with replacement, but there are no involuntary removals. At the end of year-3 teachers

[20] In an extended analysis we verify that allowing for natural teacher attrition dulls the efficiency gain from using the proportional policy (above and beyond the gain from using the global policy), although barely. The dulling effect of natural teacher attrition is so small that it is entirely hidden by sampling variance in Tables 2 and 4. Substantively, the efficiency gain from proportionality is small and similar with and without allowing for natural teacher attrition.

with three years of data are ranked and the bottom 10 percent are targeted for removal. Teachers who replaced natural exiters at the end of years one and two are excluded from the removal program. Table 6 shows our results – both policies continue to have a large positive effect overall, and the proportional policy remains the most efficient by a small margin.[21]

Changing the Gap in q Across School Types

In Tables 7 and 8 we lower the gap in average teacher quality across school types to 0.05 and 0.03, respectively. These tables are otherwise comparable to Table 2. The smaller quality gaps across school types are closer in magnitude to the gaps reported by Sass et al. (2012) between high- and low-poverty schools.[22]

The efficiency gain from the proportional policy is slightly smaller in Tables 7 and 8 relative to Table 2. In fact, in Table 8 the efficiency gain from proportionality is so small that it is not distinguishable, even to the fourth decimal place. In Appendix Tables A.1 and A.2 we show additional results using the smaller gaps where we (1) allow for natural attrition, and (2) use quality estimates based on three years of data to make the removal decisions. Building these two features into the simulations generates slightly larger efficiency gains, but the gains remain small. Returning to the logic in Figure 1, the efficiency gain is bounded from below by zero, and will reach this bound in expectation when there is no gap in teacher quality across school types.

[21] Note that there are fewer total removals in Table 6 relative to Table 4 because the removal policies are implemented conditional on teachers who have three years of data. All else equal, this reduces the achievement effect for both policies; however, the loss caused by there being fewer removals in Table 6 is more than offset by the gains that come from the improved precision of using three years of data to estimate teacher quality.

[22] These smaller gaps are also consistent with gaps in teacher quality between FRL and non-FRL students documented by Isenberg et al. (2013). Although Isenberg et al. (2013) focus on quality gaps between advantaged and disadvantaged *students*, they also report the between-school share of these gaps – although the mapping of the estimates from their study to our application is imperfect, the smaller teacher-quality gaps that we consider in this section are roughly in line with what they report.

Policy Permanency

The above results all depend on "single-shot" removal policies. It is intuitive that the efficiency gain from the proportional policy will be amplified if the policy is iterated over time (similarly to Winters and Cowen, 2013). We show the increased efficiency gains when the policies iterate in Table 9 and Appendix Tables A.3 and A.4 for scenarios where the average gap in quality across school types is 0.10, 0.05 and 0.03, respectively.

We show results based on five iterations of the policies, with removals based on single-year quality estimates and with natural attrition. The increased gains from iterating the policies can be seen by comparing the results in Table 9 to those in Table 4. The first thing to notice is that the iterative global and proportional policies have much larger achievement effects than their "single shot" counterparts. For example, the achievement gain under the global policy jumps from an average gain per student of 0.0095 (Table 4) standard deviations of student test scores based on a single year to 0.0343 (Table 9) standard deviations after iterating for five years; under the proportional policy the gain jumps from 0.0098 to 0.0350. The higher achievement gains are realized despite a smaller average gain in quality per replacement for the iterative policies. The average gain in quality per replacement declines because the iterative policies increase workforce quality over time, which lowers the per-removal gain in q in later years.

Comparing the global and proportional policies, the efficiency gains from proportionality are larger when the policies iterate. The average system-wide gain in achievement from proportionality in Table 4 is 0.0003 standard deviations of student test scores (0.0098-0.0095), while in Table 9 the average gain is 0.0007 (0.0350 – 0.0343).

Allowing for Broader Distributional Differences in q Across School Types

Thus far we have restricted the differences in the distributions of teacher quality across school types to be entirely contained by the first moment. However, Sass et al. (2012) examine distributional differences in teacher quality in Florida and North Carolina and find some evidence, particularly in Florida, to suggest that high-poverty schools have a more heterogeneous workforce as well. The wider variance of *estimated* teacher quality at high-poverty schools is likely to be partly driven by the fact that it is harder to predict student achievement for disadvantaged students, which may inflate the estimated variance of teacher quality for these students (Herrmann et al., 2013; Stacy et al., 2012).[23] Nonetheless, we take the Sass et al. estimates at face value and calibrate the data generating process around the distributional differences in teacher quality between low- and high-poverty schools that they report for Florida and North Carolina. We use the gaps based on math value-added for the calibration.

Sass et al. report teacher quality at the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the distribution for low- and high-poverty schools, respectively (see Table 6 in their paper). In Florida they estimate the gap between low- and high-poverty schools at the median to be 0.023 standard deviations of student test scores; in North Carolina the gap at the median is 0.026. The variance of teacher quality is estimated to be higher in high-poverty schools in both states. In Florida, teachers in the lower tail of the distribution in low-poverty schools appear to be markedly better than their counterparts in high-poverty schools (the gap is approximately 0.064 standard deviations at the 10^{th} percentile), while upper tail teachers in low-poverty schools are actually worse than upper tail teachers in high-poverty schools (the gap is approximately -0.021 at the 90^{th} percentile). Sass et al. also estimate that there is more

[23] The concern is that the weaker predictive power of the model for disadvantaged students creates excess residual variance for these students. Some of this residual variance may be absorbed by the estimated teacher effects, particularly with small teacher-level sample sizes, which would artificially inflate the estimated variance of teacher quality.

variance in teacher quality at high-poverty schools in North Carolina; however, the variance gap is smaller (the gaps at the 10th and 90th percentiles are 0.036 and 0.014, respectively).

Tables 10 and 11 show our results based on the Florida and North Carolina calibrations, respectively. To perform the calibrations, we first specify the distributions of *q* for teachers in type-A and type-B schools to be normal with a mean of zero and standard deviation of 0.15. Then we modify the quality estimates throughout the distribution for type-A teachers to generate the distribution-wide gaps.[24] The removal policies in the tables are based on estimated teacher quality. The best comparison table for these results is Table 8, which shows results when there are similar average differences in teacher quality to those in Tables 10 and 11, but where the differences across school types are generated entirely by shifting the distributions at the mean.

Unlike in the preceding analysis, it is not certain that the proportional policy will be the most efficient in Tables 10 and 11. On the one hand, the marginal-removal intuition from above continues to work in favor of proportionality. However, on the other hand, in both calibrations the variance of teacher quality is wider within type-B schools (particularly in Florida). All else equal, the larger variance in type-B schools increases the spread between removed teachers and the expected quality of their replacements, which makes a higher removal rate at type-B schools more desirable and pushes in favor of the global policy. Indeed, based on the Florida calibration, where teachers in high-poverty schools are parameterized to be *more effective* at the top of the distribution (Table 9), the proportional policy is slightly less efficient overall. Using the North Carolina calibration, the efficiency gain from proportionality is effectively zero.

[24] We use the reported gaps from Sass et al. at the points in the distribution for which they report the gaps directly. We fill in the gaps throughout the rest of the distribution by interpolating linearly between the points in the distribution for which the gaps are reported. For example, if they report the gap at the 50th percentile to be *p* and the gap at the 25th percentile to be *q*, we estimate the gap at the 45th percentile as [*p* + ((*q*-*p*)*(5/25))]. We hold the gaps estimated at the 10th and 90th percentiles fixed going further into the tails to help avoid overstating the distributional differences across school types.

Tables 10 and 11 illustrate that if the variance of teacher quality in high-poverty schools is higher than in low-poverty schools, the proportional policy need not be the most efficient. Still, based on our calibration following available distributional estimates, and acknowledging that these estimates may overstate the variance in teacher quality at high-poverty schools (Herrmann et al., 2013; Stacy et al., 2012), the proportional policy does not perform meaningfully worse than the global policy.

Equity

Noting that type-B schools are meant to represent high-poverty schools in our study, it may be of concern that the previously-documented efficiency gains from the proportional removal policy come through an increase in achievement at type-A schools that more than offsets a loss at type-B schools. The mechanism for the redistribution is straightforward: each removal in the general range of the distribution of q where the removals are occurring, for either school type, has a positive effect on achievement in expectation, and removals are being shifted away from type-B schools under the proportional policy. Put differently, from the perspective of the numerical simulations, the issue is that the 10-percent removal rate is a binding constraint in the achievement-maximization function. Conditional on requiring that only 10 percent of teachers be removed, the largest increase in performance at type-B schools can be achieved by removing 20 percent of teachers at these schools and no teachers at type-A schools.

A useful way to clarify the mechanism that underlies the reduced achievement gains for type-B schools under the proportional policy is to consider the case where the removal threshold is set to 50 percent. With a 50-percent removal plan, the proportional policy improves both equity *and* efficiency relative to the global policy. The reason is that the global rankings identify some teachers for removal in type-B schools who are above the 50^{th} percentile in the type-B distribution. These teachers are still below the 50^{th} percentile in the overall distribution of teacher quality, which is why they are targeted for

removal, but the expected quality of their replacements is lower than their own quality. We do not report results from a 50-percent removal policy for brevity (results available from the authors upon request); however, the 50-percent policy is a useful thought experiment to highlight the mechanism for the adverse equity effects we have documented thus far. It is an open policy question – beyond the scope of the present study – as to the equity, political and other considerations associated with how to distribute removals across school types given a fixed removal rate across the system.

A second equity consideration relates to case of a symmetric policy that awards retention bonuses to highly-effective teachers. The efficiency gain from proportionality in the retention-bonus context is accompanied by an *improvement* in performance at type-B schools (at the expense of type-A schools). To see this, consider the case of a symmetric retention bonus targeted at the top 10 percent of teachers. Based on the global teacher rankings and with perfect information, teachers in type-A schools will be overrepresented among bonus recipients to the same degree that they are underrepresented in the global removal policy in Table 1. Therefore, the proportional policy shifts retention bonuses toward type-B schools. As long as the retention bonuses have some behavioral effect, overall efficiency is improved under the proportional retention-bonus policy by logic analogous to the "marginal-removal" logic discussed in Section 4.1. In addition, the absolute number of retentions for highly effective teachers in type-B schools increases, which increases achievement at these schools.[25]

A third equity issue relates to the dynamic incentives imbedded in the proportional policy that encourage teachers to move from type-A to type-B schools. In short, in an environment where

[25] We do not formally evaluate retention policies for several reasons. Most importantly, to properly parameterize a retention bonus policy we would need to have better information about how a retention bonus might be structured to be effective and some idea of its behavioral effect. We are not aware of research evidence that provides this information (for an example of a study of an ineffective bonus program, see Clotfelter et al., 2008). The issue is that a retention bonus will not ensure retention in the same way that a determination to remove a teacher will (mostly) ensure removal.

comparisons are proportional and performance matters, teachers will be encouraged to shift to type-B schools if they are more likely to be rewarded in those schools. Ehlert et al. (2013a) also make this point.

Finally, staying within the context of our simulated removal policies, a last point on equity is that teacher turnover has a negative effect on achievement and turnover costs have not been built into the simulations thus far. The proportional removal policy shifts turnover away from type-B schools and toward type-A schools, which in the presence of a turnover-driven achievement penalty will partly offset any adverse equity effects. Furthermore, evidence from Ronfeldt et al. (2011) suggests that turnover is more costly at high-poverty schools. The asymmetric costs imply an additional mechanism by which the proportional removal policy will improve efficiency.

Tables 12 reports results from a single-year removal policy with turnover costs at type-A and type-B schools parameterized based on estimates from Ronfeldt et al. (2011).[26] The quality gap across school types is set to 0.10, per most of the preceding analysis, and removals are based on estimated quality while allowing for natural teacher attrition – the baseline comparison table without turnover costs is Table 4. Turnover costs are applied to turnovers caused by both natural and policy-based teacher attrition. The turnover costs associated with natural attrition are built into the quality estimates in row 1 of the table (in the absence of the policy).

Of course, turnover costs lower the gains from the removal policies because both policies increase turnover. Still, even with the turnover penalties in place, Table 12 shows that the policies meaningfully improve workforce quality. Moving from the global to proportional policies, the loss incurred by type-B schools from the reduced number of removals is offset by the fact that turnover costs are lower. Proportionality is again efficiency enhancing in Table 12, but unlike in the previous tables, the total efficiency gain does not come at the expense of achievement in type-B schools.

[26] Based on Ronfeldt et al. (2011), we parameterize the effect of turnover on achievement at low-poverty (type-A) and high-poverty (type-B) schools to be -0.045 and -0.075, respectively.

Table 13 expands on Table 12 by allowing the policies to iterate for 5 years. Analogous results to those shown in Table 13, but without the turnover penalty, are shown in Table 9. Table 13 is more realistic than Table 12 because it incorporates the fact that the turnover costs are single-year costs, while the benefits of re-shaping the workforce each year carry some permanency (subject to teacher attrition). The results in Table 13 show that both policies improve achievement, and that the proportional policy is most efficient. The equity tradeoff re-emerges in Table 13, but the loss for type-B schools in moving to the proportional policy is muted by the reduced turnover costs that come with the move.

6. Linking Teacher Performance in K-12 Schools to Replacement Teachers

Our analysis hinges on the link between observed- and replacement-teacher quality in type-A and type-B schools. To examine the plausibility of this link it is useful to list potential mechanisms that might drive previously-documented gaps in observed teacher quality across school types. We consider the following four possibilities:

1. The gaps reflect differences in applicant-pool quality across school types, as it has been well-established that disadvantaged schools face challenges in recruitment (Boyd et al., 2005; Clotfelter et al., 2006; Jacob, 2007; Reininger, 2012).
2. The gaps reflect differences in the quality of leadership across school types (Koedel et al., 2011). This explanation requires leadership quality to influence measured teacher quality. If this were the only source of the gaps, it could be that the applicant pools across school types are the same, but upon arrival, teachers in low-poverty schools get more support, which allows them to be more effective in the classroom.
3. The gaps reflect differences in access to instructional strategies for teachers across school types (Ehlert et al., 2013a; Raudenbusch and Willms, 1995). If better strategies are available at low-poverty schools and teachers can leverage better strategies to improve effectiveness, this could explain the gaps.

4. The gaps could reflect bias in estimation (e.g., from student sorting), in which case they would not be real.[27]

Beyond these explanations, there are undoubtedly others. However, the central feature shared by all of these potential mechanisms – and by other less-likely explanations not listed above – is that they all imply a direct connection between current teachers and their replacements. The applicant-pool explanation lends itself most directly to the way we have described the simulations above. However, it is important to recognize that the "at work" mechanisms, like leadership quality and access to instructional strategies, would be parameterized in exactly the same way. For example, if the observed gaps in teacher quality are driven entirely by gaps in leadership quality, it is still the case that replacement teachers will exhibit the same gaps as incumbents. In fact, it serves no purpose to re-run simulations tailored to these different explanations for the gap because the setup will remain the same in all cases. For the leadership example, we could just re-label the gaps in quality for replacement teachers as gaps in "access to effective principals upon entry," and the empirical results will be identical.

In summary, we cannot think of any mechanism by which teachers in high-poverty schools are less effective than their low-poverty counterparts, but where a more-effective set of replacements waits on the outside.[28] Thus, our efficiency findings are robust to the variety of potential mechanisms that might drive existing quality gaps between teachers in high- and low-poverty schools.

7. Non-Test Based Measures

Our efficiency findings will apply to any performance-based measure ranging from classroom observations to student surveys to value-added measures. The reason is that all of the mechanisms that can explain systematic differences in teacher ratings by these metrics across schooling contexts imply

[27] We ignore this fourth possibility in the discussion below – if this is the reason for the observed gaps in teacher quality across school types, the case in favor of proportionality becomes much simpler.

[28] At least in the absence of a policy designed to address the fundamental source of the observed quality gap – in the case where gap reflects differences in applicant-pool quality, an example of such a policy would be a compensating wage differential for teachers working in challenging environments.

that the differences will persist for replacement teachers as well. The logic underlying the framework for our test-based simulations translates directly to other measures of teacher performance.[29]

8. Other Considerations that Favor the Use of Proportional Models

As a practical matter, under plausible estimation conditions our results indicate that the efficiency gain from imposing proportionality will be small. In fact, in cases where there is more real variation in teacher quality in high-poverty schools, proportional policies can be slightly less efficient. However, even if we interpret our findings to indicate that the efficiency gains from proportionality are effectively zero, it is important to recognize that as long as proportionality does not have significant efficiency *costs*, it may be appealing to policymakers for a number of other reasons. Ehlert et al. (2013a) evaluate the merits of proportional policies in detail with an emphasis on school-level evaluations. The insights from their study carry over to teacher evaluations as well. In short, these authors advocate that proportional evaluations are desirable because they: (1) generate performance signals that will be useful for improving instruction in K-12 schools, (2) elicit optimal effort from teachers, and (3) avoid exacerbating well-documented inequities in the labor markets faced by advantaged and disadvantaged schools.

We avoid a lengthy review of the arguments in Ehlert et al. (2013a) in this paper. Of importance for our study is that Ehlert et al. identify several substantive reasons for policymakers to prefer a proportional evaluation system as long as there are not mitigating negative consequences. Here, we

[29] Teacher performance measures need not derive from models of student achievement to be constructed as proportional. As noted above, proportionality can always be achieved via *ex post* regression adjustment at the teacher level.

dispel one possible concern by showing that proportional policies do not have efficiency costs, and under plausible conditions they can be efficiency enhancing.[30]

We also note an additional benefit of proportionality not covered by Ehlert et al. (2013a): as teacher evaluation systems come online at scale, concerns about fairness are increasingly common. As just one example, the Boston Globe recently reported on the Boston teacher union's concern over the school district's evaluation system. Black and Hispanic teachers in Boston are significantly more likely to be identified as underperforming relative to white teachers (Vaznis, 2013). Specifically, the article reports that based on the current evaluation system, black teachers in Boston are three times more likely than white teachers to be placed on a "direct growth plan" or "improvement plan" – both plans can lead to termination. The president of the teacher's union, Richard Stutman, is quoted as saying: "I don't know how [the School Department] can defend a system that is disproportionately identifying black and Hispanic teachers."

One factor that may contribute to the racial differences in performance ratings across teachers in Boston is differences in teaching circumstance. For example, schools where the student body is disproportionally African American also likely have a disproportionate share of African American teachers, and lower achievement growth (e.g., see Dee, 2004). A proportional model can help mitigate differences in teacher assessments that fall along this and other contextual lines. In fact, if desired, proportionality could be explicitly imposed at the teacher level so as to guard against disproportionate identification of certain types of teachers as high- and/or low-performing.[31]

[30] Again, note that some of the benefits of proportionality discussed by Ehlert et al. (2013a) may improve efficiency by affecting educator behavior. Any efficiency gains caused by educators' behavioral responses to proportional evaluations are not reflected in our analysis (because we do not have enough information to parameterize teachers' behavioral responses and their effects on achievement). Therefore, we may understate the efficiency gains from proportional evaluations.

[31] We do not advocate for or against the use of proportionality in this way – we simply note that imposing teacher-level proportionality would be consistent with stated objectives in many districts to increase (or not decrease) workforce diversity. In Boston, for example, the city council and city youth organizations are pressuring the district

While it is outside of the scope of this paper to formally evaluate the costs and benefits of implementing proportional evaluations for this purpose, the fairness issue in Boston is one that proportional evaluations can help to address. Our results, which show that proportional evaluations can be used to resolve real-world problems with teacher evaluations without efficiency costs, will be of value to educational decision makers in this era where rigorous teacher evaluation systems are increasingly common.

9. Concluding Remarks

Many state and local education agencies have developed, or are in the process of developing, rigorous teacher evaluation systems. An impetus for these systems is the consistent finding in the research literature that there is considerable variation in teacher quality and that access to effective teaching meaningfully affects students' immediate and longer-term outcomes (Hanushek and Rivkin, 2010; Chetty, Friedman and Rockoff, 2011). However, despite the rapid growth in the development of educator evaluation systems, many of the design details surrounding these systems remain unresolved.

The contribution of the present study is to examine the efficiency implications of imposing proportionality in teacher evaluation systems. We show that under plausible conditions – most notably when the gaps in observed teacher performance across schooling contexts carry over for replacement teachers – proportional evaluations can be efficiency enhancing. While the efficiency gains that we document under real-world evaluation conditions are small; in conjunction with the other benefits that proportionality offers, and the potential for proportionality to improve efficiency along other dimensions that we do not consider (most notably by positively affecting educator behavior – see Ehlert

to increase the diversity of the teaching workforce (Vaznis, 2013). Even in the absence of imposing explicit proportionality using teacher characteristics, controlling for school context more generally should help to mitigate problems like the one in Boston because of the correlation between the shares of students and teachers by race within schools (also, Grissom and Keiser (2011) show that there is a similar racial correlation for school principals).

et al., 2013a), our findings point to proportional evaluations as being a viable alternative for educational administrators charged with developing and implementing teacher evaluation systems.

References

Arcaira, Erikson, Beatrice Birman, Stephen Coleman, Erin Dunlop, Michael Feuer, Maxine Freund, Steve Glazerman, Jane Hannaway, Heather Harding, Jaclyn MacFarlane, Taunya Nesin, Umut Ozek, Andrea Palmiter, Brenda Turnbull and Elias Walsh. 2013. Evaluation of the DC Public Education Reform Amendment Act (PERAA): Report No. 2, School Year 2011-2012. Report Published by the Education Consortium for Research and Evaluation at The George Washington University.

Barlevy, Gary and Derek Neal. 2012. Pay for Percentile. *American Economic Review* 102(5), 1805-31.

Betebenner, Damian W. 2009. Norm-and Criterion-Referenced Student Growth. Educational Measurement: Issues and Practice 28(4), 42-51.

Boyd, Donald, Hamilton Lankford, Susanna Loeb. 2005. The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. *Journal of Policy Analysis and Management* 24(1), 113-132.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. 2012. Measuring Test Measurement Error: A General Approach. NBER Working Paper No. 18010.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2011. The Long-Term Impacts of Teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper No. 17699.

Chute, Eleanor. 2013. New Teacher Evaluation Process Set to Begin in Pittsburgh Public Schools. *Pittsburgh Post-Gazette* (08.13.2013).

Bill and Melinda Gates Foundation. 2013. Culminating Findings from the MET Project's Three-Year Study. Policy Report.

Clotfelter, Charles, Helen F. Ladd, Jacob Vigdor and Justin Wheeler. 2006. High-Poverty Schools and the Distribution of Teachers and Principals. *North Carolina Law Review* 85, 1345-1379.

Clotfelter, Charles T., Elizabeth J. Glennie, Helen F. Ladd, and Jacob L. Vigdor. 2008. Teacher Bonuses and Teacher Retention in Low-Performing Schools. *Public Finance Review* 36(1), 63-87.

Dee, Thomas. 2004. Teachers, Race and Student Achievement. *Review of Economics and Statistics* 86(1), 195-210.

Dee, Thomas and James Wyckoff. 2013. Incentives, Selection and Teacher Performance. Evidence from IMPACT. NBER Working Paper No. 19529.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2013a. Selecting Growth Measures for School and Teacher Evaluations: Should Proportionality Matter? CALDER Working Paper No. 80.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2013b. The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence from School- and Teacher-Level Models in Missouri. *Statistics and Public Policy* 1(1): 19-27.

Feng, Li and Tim R. Sass. 2011. Teacher Quality and Teacher Mobility. CALDER Working Paper No. 57.

Goldhaber, Dan and Duncan Chaplin. 2012. Assessing the "Rothstein Falsification Test." Does it Really Show Teacher Value-added Models are Biased? CEDR Working Paper No. 2012-1.

Goldhaber, Dan, Pete Goldschmidt and Fannie Tseng. 2013. Teacher Value-Added at the High-School Level: Different Models, Different Answers? *Educational Evaluation and Policy Analysis* 35(2), 220-236.

Goldhaber, Dan and Michael Hansen (forthcoming). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. *Economica*.

Goldhaber, Dan, Joe Walch and Brian Gabele. 2013. Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. *Statistics and Public Policy* 1(1): 28-39.

Grissom, Jason A., and Lael Keiser. 2011. A Supervisor Like Me: Race, Representation, and the Satisfaction and Turnover Decisions of Public Sector Employees. *Journal of Policy Analysis and Management* 30(3): 557-580.

Hanushek, Eric A and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100(2), 267-271.

Sass, Tim R., Anastasia Semykina and Douglas N. Harris. 2014. Value-Added Models and the Measurement of Teacher Productivity. *Economics of Education Review* 38(1), 9-23.

Herrmann, Mariesa, Elias Walsh, Eric Isenberg and Alexandra Resch. 2013. Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels. Policy Report, Mathematica Policy Research.

Isenberg, Eric, Jeffrey Max, Philip Gleason, Liz Potamites, Robert Santillano, Heinrich Hock and Michael Hansen. 2013. Access to Effective Teaching for Disadvantaged Students. Report: United States Department of Education.

Jacob Brian. 2007. The Challenges of Staffing Urban Schools with Effective Teachers. *The Future of Children* 17(1), 129-153.

Kane, Tom J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. What Does Certification Tell us about Teacher Effectiveness? Evidence from New York City. *Economics of Education Review* 27(6), 615-631.

Koedel, Cory and Julian R. Betts. 2007. Re-Examining the Role of Teacher Quality in the Educational Production Function. University of Missouri Working Paper No. 07-08.

Koedel, Cory and Julian R. Betts. 2011. Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy* 6(1), 18-42.

Koedel, Cory, Jason A. Grissom, Shawn Ni and Michael Podgursky. 2011. Pension-Induced Rigidities in the Labor Market for School Leaders. CALDER Working Paper No 62.

Koedel, Cory, Rebecca Leatherman and Eric Parsons. 2012. Test Measurement Error and Inference from Value-Added Models. *The B.E. Journal of Economic Analysis & Policy* 12(1).

McCaffrey, Daniel F., J.R. Lockwood, Daniel M. Koretz and Laura S. Hamilton. 2003. Evaluating Value-Added Models for Teacher Accountability. Santa Monica, CA: The RAND Corporation.

McCaffrey, Daniel F., Tim R. Sass, J.R. Lockwood and Kata Mihaly. 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy* 4(4), 572-606.

Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger and J.R. Lockwood. 2013. A Composite Estimator of Effective Teaching. *RAND External Publication*, EP-50155.

Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. 2004. How Large are Teacher Effects? *Educational Evaluation and Policy Analysis* 26(3), 237-257.

Polikoff, Morgan S. 2013. The Stability of Observational and Student Survey Measures of Teaching Effectiveness. Paper presented at the 2013 Annual Conference of the Association for Education Finance and Policy, New Orleans, LA.

Polikoff, Morgan S., Andrew J. McEachin, Stephani L. Wrabel and Matthew Duque. Forthcoming. The Waive of the Future? School Accountability in the Waiver Era. *Educational Researcher*.

Raudenbush, Stephen and J. Douglas Willms. 1995. The Estimation of School Effects. *Journal of Educational and Behavioral Statistics* 20(4): 307-335.

Reininger, Michelle. 2012. Hometown Disadvantage? It Depends on Where You're From: Teachers' location preferences and the implications for staffing schools. *Educational Evaluation and Policy Analysis* 34(2), 127-145.

Rivkin, Steven G., Eric A. Hanushek and John F. Kain. 2005. Teachers, Schools and Academic Achievement. *Econometrica* 73(2), 417-58.

Ronfeldt, Matthew, Hamilton Lankford, Susanna Loeb and James Wyckoff. 2011. How Teacher Turnover Harms Student Achievement. Working Paper No. 17176, National Bureau of Economic Research.

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio and Li Feng. 2012. Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools. *Journal of Urban Economics* 72, 104-122.

Schochet, Peter Z. and Hanley S. Chiang (forthcoming). What are Error Rates for Classifying Teacher and School Performance Measures Using Value-Added Models? *Journal of Educational and Behavioral Statistics*.

Schotter, Andrew and Keith Weigelt. 1992. Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results. *Quarterly Journal of Economics* 107 (2), 511-539.

Shifrer, Dara, Ruth Lopez Turley and Holly Heard. 2013. Houston Independent School District's Aspire Program: Estimated Effects of Receiving Financial Awards. Houston Educational Research Consortium Policy Report.

Stacy, Brian, Cassandra Guarino, Mark Reckase and Jeffrey Wooldridge. 2012. Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve? Unpublished manuscript.

Staiger, Douglas O. and Jonah E. Rockoff. 2010. Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.

Strunk, Katharine O., Tracey L. Weinstein and Reino Makkonnen. 2013. Sorting out the Signal:

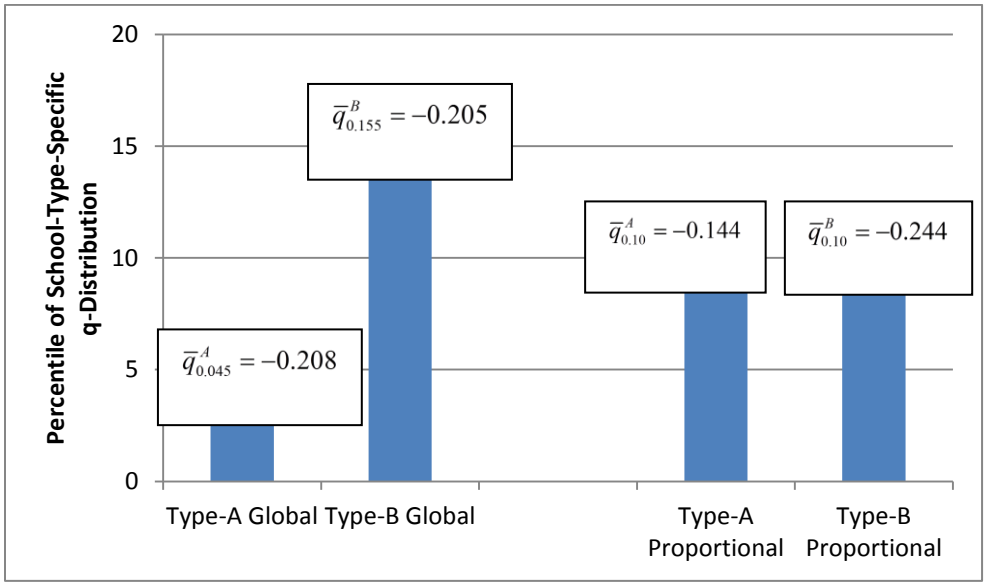
Do Multiple Measures of Teachers' Effectiveness Provide Consistent Information to Teachers and Principals? Working Paper, University of Southern California.

Vaznis, James. 2013. Union Says Teacher Evaluation Plan has Race Bias. *Boston Globe* (04.23.2013).

Winters, Marcus A. and Joshua M. Cowen. 2013. Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management* 32(3), 634-654.

Figures

Figure 1. Average Percentile Rankings and *q*-Values for Marginally Removed Teacher at Each School Type Using Global and Proportional Rankings. Based on 1,000 Simulations.



Notes: The expected value of *q* for replacement teachers in the figure is 0.05 at type-A schools and -0.05 at type-B schools. The standard deviation of *q* within each sector is 0.15.

Tables

Table 1. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Persistent Effectiveness (q), Perfectly Observed.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0500	-0.0499	0.0500	0.0499
Average Quality with Policy	0.0646	-0.0144	0.0762	-0.0236
Achievement Gain	0.0146	0.0355	0.0262	0.0263
Combined Achievement Gain (weighted)	0.0250		0.0263	
Number of Teachers Removed	13.894	46.106	30	30
Average Gain in q per Replacement	0.315	0.231	0.262	0.263
Average Gain in q per Replacement (weighted)	0.250		0.263	

Notes: Average values across 1,000 iterations of the simulated policy are reported. Achievement gains are equal to 10 percent of the average gain in q per removed teacher because 10 percent of the workforce – and therefore 10 percent of students (per the homogenous class size built into the simulations) – are affected.

Table 2. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$).

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0498	-0.0499	0.0498	-0.0499
Average Quality with Policy	0.0598	-0.0345	0.0627	-0.0370
Achievement Gain	0.0100	0.0153	0.0129	0.0129
Combined Achievement Gain (weighted)	0.0127		0.0129	
Number of Teachers Removed	21.616	38.384	30	30
Average Gain in q per Replacement	0.139	0.119	0.129	0.129
Average Gain in q per Replacement (weighted)	0.127		0.129	

Notes: See notes for Table 1.

Table 3. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Estimates of Persistent Effectiveness with Bias ($\hat{\gamma} + B_T$).

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0500	-0.0495	0.0500	-0.0495
Estimated Average Quality in the Absence of Policy (with Bias)	0.0700	-0.0695	0.0700	-0.0695
Average Quality with Policy	0.0589	-0.0334	0.0630	-0.0366
Achievement Gain	0.0089	0.0162	0.0130	0.0130
Combined Achievement Gain (weighted)	0.0125		0.0130	
Number of Removed Teachers	18.328	41.672	30	30
Gain in *q* per Replacement, by Type	0.146	0.116	0.130	0.130
Average Gain in *q* per Replacement (weighted)	0.125		0.130	

Notes: See notes for Table 1. See text for a discussion of how we introduce bias into the estimates of teaching effectiveness.

Table 4. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$), with Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy (but with natural attrition)	0.0494	-0.0481	0.0494	-0.0481
Average Quality with Policy	0.0571	-0.0367	0.0593	-0.0383
Achievement Gain	0.0077	0.0113	0.0099	0.0097
Combined Achievement Gain (weighted)	0.0095		0.0098	
Number of Teachers Removed	17.332	30.334	24.183	23.611
Average Gain in *q* per Replacement	0.133	0.112	0.122	0.124
Average Gain in *q* per Replacement (weighted)	0.120		0.123	

Notes: See notes for Table 1. Teacher attrition is parameterized as described in the text. Note that with teacher attrition the gain in *q* per replacement teacher is no longer equal to 10 times the gain in total achievement because fewer than 10 percent of teachers are involuntarily removed.

Table 5. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) Using 3 Years of Data to Obtain Estimates, without Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0504	-0.0499	0.0504	-0.0499
Average Quality with Policy	0.0632	-0.0265	0.0688	-0.0315
Achievement Gain	0.0127	0.0234	0.0184	0.0184
Combined Achievement Gain (weighted)	0.0181		0.0184	
Number of Teachers Removed	18.220	41.780	30	30
Average Gain in *q* per Replacement	0.2097	0.1679	0.184	0.184
Average Gain in *q* per Replacement (weighted)	0.181		0.184	

Notes: See notes for Table 1.

Table 6. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) Using 3 Years of Data to Obtain Estimates, with Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy (but with natural attrition)	0.0493	-0.0463	0.0493	-0.0463
Average Quality with Policy	0.0574	-0.0316	0.0608	-0.0348
Achievement Gain	0.0081	0.0146	0.0115	0.0115
Combined Achievement Gain (weighted)	0.0114		0.0115	
Number of Teachers Removed	12.455	27.615	20.033	19.752
Average Gain in *q* per Replacement	0.195	0.159	0.173	0.174
Average Gain in *q* per Replacement (weighted)	0.170		0.173	

Notes: See notes for Table 4.

Table 7. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with the Mean Gap in Effectiveness Across Sectors Reduced to 0.05 and Without Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0248	-0.0249	0.0248	-0.0249
Average Quality with Policy	0.0363	-0.0108	0.0376	-0.0120
Achievement Gain	0.0115	0.0141	0.0128	0.0130
Combined Achievement Gain (weighted)	0.0128		0.0129	
Number of Teachers Removed	25.817	34.183	30	30
Average Gain in *q* per Replacement	0.133	0.124	0.128	0.130
Average Gain in *q* per Replacement (weighted)	0.128		0.129	

Notes: See notes for Table 1.

Table 8. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with the Mean Gap in Effectiveness Across Sectors Reduced to 0.03 and Without Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0150	-0.0151	0.0150	-0.0151
Average Quality with Policy	0.0272	-0.0013	0.0280	-0.0021
Achievement Gain	0.0122	0.0138	0.0130	0.0130
Combined Achievement Gain (weighted)	0.0130		0.0130	
Number of Teachers Removed	27.327	32.676	30	30
Average Gain in *q* per Replacement	0.134	0.126	0.130	0.130
Average Gain in *q* per Replacement (weighted)	0.130		0.130	

Notes: See notes for Table 1.

Table 9. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with Policy Permanency. Removals are Based on Single-Year Quality Estimates and Policy Iterates for Five Consecutive Years, with Natural Teacher Attrition. Achievement Gains and Workforce Quality Improvements are Reported After Year Five.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy (but with natural attrition)	0.0487	-0.0456	0.0487	-0.0456
Average Quality with Policy	0.0789	-0.0073	0.0842	-0.0111
Achievement Gain	0.0302	0.0383	0.0355	0.0344
Combined Achievement Gain (weighted)	0.0343		0.0350	
Number of Teachers Removed	96.633	162.531	130.798	129.184
Average Gain in *q* per Replacement	0.094	0.071	0.081	0.080
Average Gain in *q* per Replacement (weighted)	0.079		0.081	

Notes: See notes for Table 4.

Table 10. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$), with Distributional Differences in Teacher Effectiveness across School Types Based on Florida Estimates from Sass. et al., Without Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0212	0	0.0212	0
Average Quality with Policy	0.0295	0.0141	0.0303	0.0129
Achievement Gain	0.0083	0.0141	0.0092	0.0130
Combined Achievement Gain (weighted)	0.0112		0.0111	
Number of Teachers Removed	26.717	33.283	30	30
Average Gain in *q* per Replacement	0.094	0.127	0.092	0.130
Average Gain in *q* per Replacement (weighted)	0.112		0.111	

Notes: See notes for Table 1.

Table 11. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$), with Distributional Differences in Teacher Effectiveness across School Types Based on North Carolina Estimates from Sass. et al., Without Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0229	0	0.0229	0
Average Quality with Policy	0.0338	0.0138	0.0347	0.0130
Achievement Gain	0.0109	0.0138	0.0118	0.0130
Combined Achievement Gain (weighted)	0.0124		0.0124	
Number of Teachers Removed	27.690	32.312	30	30
Average Gain in *q* per Replacement	0.118	0.128	0.118	0.130
Average Gain in *q* per Replacement (weighted)	0.124		0.124	

Notes: See notes for Table 1.

Table 12. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$), with Turnover Penalty and Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy (with natural attrition)	0.0410	-0.0630	0.0410	-0.0630
Average Quality with Policy	0.0467	-0.0590	0.0473	-0.0589
Achievement Gain	0.0058	0.0040	0.0063	0.0041
Combined Achievement Gain (weighted)	0.0049		0.0052	
Number of Teachers Removed	20.714	26.872	24.029	23.592
Average Gain in *q* per Replacement (with turnover penalty built in)	0.084	0.045	0.079	0.052
Average Gain in *q* per Replacement (weighted, with turnover penalty built in)	0.061		0.066	

Notes: See notes to Table 4. The turnover penalty also applies to teachers who leave voluntarily and is built into the average-quality estimates reported in row 1 as indicated in the table.

Table 13. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$), with Turnover Penalty and Natural Teacher Attrition. Removals are Based on Single-Year Quality Estimates and Policy Iterates for Five Consecutive Years. Achievement Gains and Workforce Quality Improvements are Reported After Year Five.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy (with natural attrition)	0.0448	-0.0522	0.0448	-0.0522
Average Quality with Policy	0.0707	-0.0221	0.0757	-0.0251
Achievement Gain	0.0258	0.0301	0.0309	0.0271
Combined Achievement Gain (weighted)	0.0280		0.0290	
Number of Teachers Removed	93.346	163.752	130.205	127.898
Average Gain in *q* per Replacement (with turnover penalty built in)*	0.083	0.055	0.071	0.064
Average Gain in *q* per Replacement (weighted, with turnover penalty built in)	0.065		0.067	

* The turnover penalty is built into each exit so that the results are reported in a manner consistent with the results reported in Table 12. However, note that the turnover penalty disappears after year-1 and the gain in *q* persists subject to natural attrition as described in the text.

Notes: See notes to Table 12.

Appendix A. Supplementary Tables

Appendix Table A.1. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with the Mean Gap in Effectiveness Across Sectors Reduced to 0.05, Using 3 Years of Data to Obtain Estimates, with Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0254	-0.0223	0.0254	-0.0223
Average Quality with Policy	0.0354	-0.0095	0.0372	-0.0109
Achievement Gain	0.0100	0.0128	0.0117	0.0114
Combined Achievement Gain (weighted)	0.0114		0.0115	
Number of Teachers Removed	16.075	23.890	19.912	19.751
Average Gain in *q* per Replacement	0.187	0.161	0.177	0.172
Average Gain in *q* per Replacement (weighted)	0.171		0.175	

Notes: See notes for Table 4.

Appendix Table A.2. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with the Mean Gap in Effectiveness Across Sectors Reduced to 0.03, Using 3 Years of Data to Obtain Estimates, with Natural Teacher Attrition.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0158	-0.0129	0.0158	-0.0129
Average Quality with Policy	0.0264	0.0001	0.0276	-0.0011
Achievement Gain	0.0106	0.0130	0.0118	0.0119
Combined Achievement Gain (weighted)	0.0118		0.0118	
Number of Teachers Removed	17.558	22.540	19.982	19.881
Average Gain in *q* per Replacement	0.181	0.173	0.177	0.179
Average Gain in *q* per Replacement (weighted)	0.176		0.178	

Notes: See notes for Table 4.

Table A.3. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with Policy Permanency. Removals are Based on Single-Year Quality Estimates and Policy Iterates for Five Consecutive Years, with Natural Teacher Attrition. Achievement Gains and Workforce Quality Improvements are Reported After Year Five. The Mean Gap in Effectiveness Across Sectors Reduced to 0.05.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0253	-0.0221	0.0253	-0.0221
Average Quality with Policy	0.0575	0.0152	0.0603	0.0126
Achievement Gain	0.0322	0.0372	0.0350	0.0346
Combined Achievement Gain (weighted)	0.0347		0.0348	
Number of Teachers Removed	113.064	146.793	130.582	129.481
Average Gain in *q* per Replacement	0.0854	0.0761	0.080	0.080
Average Gain in *q* per Replacement (weighted)	0.080		0.080	

Notes: See notes for Table 4.

Table A.4. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Unbiased Estimates of Persistent Effectiveness ($\hat{\gamma}$) with Policy Permanency. Removals are Based on Single-Year Quality Estimates and Policy Iterates for Five Consecutive Years, with Natural Teacher Attrition. Achievement Gains and Workforce Quality Improvements are Reported After Year Five. The Mean Gap in Effectiveness Across Sectors Reduced to 0.03.

	Removal Policy based on Global Rankings		Removal Policy based on Proportional Rankings	
	Type-A Schools	Type-B Schools	Type-A Schools	Type-B Schools
Average Quality in the Absence of Policy	0.0159	-0.0121	0.0159	-0.0121
Average Quality with Policy	0.0496	0.0238	0.0508	0.0230
Achievement Gain	0.0337	0.0359	0.0350	0.0351
Combined Achievement Gain (weighted)	0.0348		0.0350	
Number of Teachers Removed	120.454	139.366	130.277	129.611
Average Gain in *q* per Replacement	0.084	0.077	0.081	0.081
Average Gain in *q* per Replacement (weighted)	0.080		0.081	

Notes: See notes for Table 4.