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 teacher ratings to help shape the workforce and define efficiency in terms of student achievement. Our analysis shows 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 evaluations offer a number of other policy benefits.

JEL Codes: I20, J48

Acknowledgments
Koedel is in the department of economics and Truman School of Public Affairs, and Li is in the department of economics, at the University of Missouri, Columbia. The authors thank Eric Isenberg, Dave Kaplan, Eric Parsons and participants at the 2014 CALDER and AEFP conferences for useful comments. They gratefully acknowledge research support from CALDER. The usual disclaimers apply.
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.\textsuperscript{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 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 define efficiency in terms of student achievement – the most efficient policy is the one that results in the highest achievement in total. We compare evaluation systems that rank order teachers based on (1) proportional estimates of teacher quality, which force comparisons to be between equally-circumstanced teachers, and (2) global estimates of quality that compare teachers to each other regardless of teaching circumstance.\textsuperscript{2} The context for the comparison is a removal policy targeted at the bottom 10 percent of teachers. Our analysis is performed using simulated data, which we construct 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.,

\textsuperscript{1} A number of states have enacted legislation mandating performance-based evaluations and high stakes have been attached in some cases. Senate Bill 736 in Florida (2011) and House Bill 1001 in Colorado (2012) are examples of such legislation. Similar legislation is being considered and/or implemented in other states, including Michigan and Pennsylvania. Some large school districts are also independently developing performance-based teacher evaluations. The Los Angeles Unified School District (Strunk, Weinstein and Makkonen, 2014) and Washington DC Public School District (Arcaira et al., 2013) are examples.

\textsuperscript{2} Ehler et al. (forthcoming) discuss proportional evaluations in detail. The proportionality principle – although it is referred to by a variety of different names – has been examined in a number of different contexts (e.g., see Barlevy and Neal, 2012; Calsamiglia, Franke and Rey-Biel, 2014; Schotter and Weigelt, 1992).
classroom observations, student evaluations). However, the substance of our findings will apply to any 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 (Mihaly et al., 2013).

To illustrate how proportional and global teacher 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. A quality-based removal policy that depends on global rankings will identify more teachers from type-B schools to be removed. In contrast, an analogous policy based on proportional rankings, which force equally-circumstanced comparisons, will ensure that an equal number of teachers from type-A and type-B schools are removed.

It is straightforward to show that the proportional policy is more efficient 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, Walsh and Gabele, 2013; Isenberg et al., 2013; Sass et al., 2012). The key insight underlying the efficiency gain is that the effect of a targeted removal policy depends not only on the quality of the teachers being removed, but also on the quality of replacements. 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 replacement teachers at these schools as well. After taking direct account of the link between observed teacher quality and replacement teacher quality for schools in different contexts, we show that the proportional policy is the most efficient in terms of raising total student achievement.

---

3 There is a vast literature examining test-based performance measures and their properties — examples of studies include Chetty, Friedman and Rockoff (2014a), Ehlert et al. (forthcoming, 2013), Goldhaber, Walsh and Gabele (2013) and Sass, Semykina and Harris (2014). Hanushek and Rivkin (2010) provide an overview of the test-based literature. Researchers are just beginning to rigorously explore the properties of non-test based measures (Kane and Staiger, 2012).

4 Type-A schools can be thought of as low-poverty schools and type-B schools as high-poverty schools. This example is motivated by evidence showing gaps in teacher quality between high- and low-poverty schools (e.g., see Arcaira et al., 2013; Goldhaber, Walsh and Gabele, 2013; Isenberg et al., 2013; Sass et al., 2012). There are also a number of studies that discuss general recruiting challenges for high-poverty schools (e.g., see Boyd et al., 2005; Clotfelter et al., 2006).
Although there is a strong efficiency rationale for proportionality, we find that in real-world applications the efficiency gain from imposing proportionality will be small. One reason is that teachers are evaluated in practice using imprecise measures, which attenuates the efficiency gain. Another reason is that the efficiency gain 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 does not meaningfully underperform the global policy in our study, it may be preferable for several reasons. One issue is that our research design focuses on a single mechanism by which proportional evaluations affect efficiency – namely, how they influence which teachers are removed and replaced. We do not allow the proportional policy to generate efficiency gains for other reasons, although studies in other contexts suggest efficiency gains for other reasons are likely (e.g., Calsamiglia, Franke and Rey-Biel, 2014; Schotter and Weigelt, 1992). Proportional evaluations also offer several other benefits that are not directly tied to the efficient production of student achievement, which we review briefly below. In summary, our finding that proportional evaluations do not have efficiency costs along the removal-replacement dimension, and under some conditions represent an efficiency improvement, points toward proportionality being a viable design feature of teacher evaluation systems.

2. Background

2.1 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 (2014b) 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 to observable characteristics, 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.

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. Winters and Cowen (2013) provide 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. (Dee and Wyckoff, 2013; Mihaly et al., 2013; Strunk, Weinstein and Makkonen, 2014). 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. Our study aims to inform the decision-making process by providing concrete evidence about a particular aspect of model selection – namely, we examine the efficiency effects of imposing proportionality in teacher evaluations.

2.2 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., 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. (forthcoming) estimate a proportional model for schools in Missouri using a two-step fixed effects

\[ \text{Legislation in a number of states mandates that a minimum percentage of teachers’ overall ratings depend on student achievement growth. Examples of states with such mandates include Florida and Colorado.} \]
procedure. The key feature of their approach is that it partials out the variance in student test scores attributable to observable characteristics of students and schools prior to estimating school value-added. Proportionality can also be enforced outside of the model of student achievement – for example, by performing ex post regression adjustments at the unit of evaluation (e.g., school, teacher). We refer interested readers to Ehlert et al. (forthcoming) for details.\(^6\)

Enforcing the proportionality property in teacher rankings can have meaningful evaluative consequences when teacher quality differs systematically across different types of schools, and available research shows that estimated teacher quality is consistently lower in high-poverty schools (Goldhaber, Walch and Gabele, 2013; Isenberg et al., 2013; and Sass et al., 2012). We take the previously-documented gap in estimated teacher quality between high- and low-poverty schools as given and consider cases where (1) the gap reflects a true 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, which is the best-case scenario for the use of global rankings.\(^7\)

2.3 Non-Efficiency Considerations

The purpose of the analysis that follows is to examine the efficiency implications of proportionality. However, first we briefly review other, non-efficiency-based considerations related to the use of proportional evaluations for teachers. Ehlert et al. (forthcoming) discuss the policy merits of proportionality in detail. They argue that proportional evaluations are desirable for the following

---

6 One might think that using models that include school fixed effects would ensure proportionality, but this is only true if teachers are entirely nested within schools (e.g., in a single-year model where teachers do not change schools and the teacher effects are constrained to sum to zero within schools). In school-fixed-effects models that rely on school switchers for identification, teachers are still compared across schools and thus proportional rankings need not result (for a detailed discussion on the identification of teacher effects using school switchers in models with school fixed effects, see Mansfield, forthcoming). As a practical matter, we are not aware of any teacher evaluation systems used in practice that rely on a model with school fixed effects.

7 If the estimated teacher-quality gap between high- and low-poverty schools is driven by inadequate controls then the proportional model would be preferred for other reasons beyond those discussed in this paper – namely, bias reduction.
reasons: (1) they generate performance signals that are useful for improving instruction in K-12 schools, (2) they elicit optimal effort from teachers, and (3) they 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. (forthcoming) here. Of importance is that Ehlert et al. identify several substantive reasons for policymakers to prefer a proportional evaluation system as long as there are no mitigating negative consequences (e.g., efficiency costs).\(^8\)

We also note an additional benefit of proportionality not covered by Ehlert et al. (forthcoming). As teacher evaluation systems come online at scale, concerns about fairness are increasingly common, and proportionality can be used to help address these concerns. 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. Specifically, Vaznis (2013) 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. Richard Stutman, the president of the teacher’s union, 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 disproportionately 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

---

\(^8\) Again, some of the benefits of proportionality discussed by Ehlert et al. (forthcoming) may improve efficiency by affecting educator behavior (also see Calsamiglia, Franke and Rey-Biel, 2014; Schotter and Weigelt, 1992). Any efficiency gains caused by educators’ behavioral responses to proportionality are not reflected in our analysis and we may understate the efficiency gains from proportional evaluations for this reason.
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. While it is outside of the scope of our study 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 if desired.⁹

3. The Efficiency Rationale for Proportional Teacher Evaluations

3.1 Basic Theoretical Framework

In this section we illustrate the efficiency rationale for proportionality with a simple model where again, there are two types of schools: (1) type-A, low-poverty schools, and (2) type-B, high-poverty schools. We incorporate the empirical regularity that in the absence of forced proportional comparisons, teacher quality in low-poverty schools is higher on average than teacher quality in high-poverty schools. Note that we could also set up the model so that schools differ by a continuous poverty measure, such as the share of students eligible for free/reduced-price lunch or the share of disadvantaged-minority students. We use type-A and type-B schools for ease of presentation and without loss of generality.

In the model, there are an equal number of teachers in each sector and teachers are drawn from the following type-specific, uniform quality distributions:

\[ q_A \sim U(a_A, b_A) \]
\[ q_B \sim U(a_B, b_B) \]  

(1)

⁹ 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, Vaznis (2013) reports that the city council and city youth organizations are pressuring the district to increase the diversity of the teaching workforce. For a general discussion of fairness issues in teacher evaluations see Polikoff et al. (2014).
where $b_A - a_A = b_B - a_B$. Thus, there is a gap in average teacher quality across sectors but the variance of teacher quality is the same within each sector (we will consider scenarios with unequal variance later on). Teacher quality can be mapped directly to achievement – i.e., $a_A, a_B, b_A$ and $b_B$ can be defined in terms of teacher effects on achievement.

We examine a hypothetical policy that aims to increase student achievement by replacing ineffective teachers with new teachers, where new teachers are drawn from the same type-specific distributions shown in (1) and the quality of incumbent teachers is known. The achievement maximization problem can be written as follows:

$$
MAX_{R_A, R_B} E(Q) = (\mu_A R_A - S_{A,R_A}) + (\mu_B R_B - S_{B,R_B})
$$

(2)

where $\mu_A$ and $\mu_B$ represent the expected quality of replacement teachers and $R_A$ and $R_B$ are the numbers of removed/replacement teachers in each sector, respectively. $S_{j,R_j}$ measures the summative quality of removed teachers in sector $j$, which depends on $R_j$. Specifically, given that teacher quality for incumbents is known, $S_{j,R_j}$ is equal to the expected sum of the first $R_j$ order statistics drawn from the distribution in sector $j$ and can be expressed as:

$$
S_{j,R_j} = E(\sum_{k=1}^{R_j} q_{jk}) = \sum_{k=1}^{R_j} E(q_{jk})
$$

(3)

where $q_{jk}$ is the $kth$ smallest value of teacher quality in the sector-$j$ distribution.

Consider the constrained optimization problem where the objective is to maximize equation (2) subject to a constraint dictating the number of removals, $R_A + R_B = \bar{R}$. Intuitively, optimization will occur at the point where the value of the expected gain from the marginal removal is equated across sectors; i.e., where $\mu_A - r_A^m = \mu_B - r_B^m$, with $r_j^m$ equal to the quality of the marginally removed teacher in sector $j$. If the gain associated with the marginal removal across sectors is not equated,
expected achievement can be increased by changing the sector from which the marginal removal is
drawn. The global policy fails to maximize achievement because removals under the global policy do
not fully account for context. Thus, the global policy is designed to equate \( r_A^m \) and \( r_B^m \). However,
because \( \mu_A \neq \mu_B \), forcing the quality of the marginal removal across sectors to be equal ensures a gap
in the *quality gain* from the marginal replacement, and thus inefficiency. In contrast, by permitting
removals such that \( r_A^m \neq r_B^m \), the proportional policy facilitates efficient removals.

Formally, we can solve the model to show that proportionality satisfies the conditions for a
maximum. We begin by unpacking equation (3). First, note that by the probability integral transformation, for any random variable \( X \) from a continuous distribution with cumulative
distribution function (CDF) \( F_X \), the random variable \( Z = F_X(X) \) follows a standard uniform
distribution. Put differently, drawing randomly from the CDF of any continuous distribution
(including a uniform distribution as in our context) is equivalent to drawing from a standard uniform
distribution – i.e., \( F(X_k) \overset{d}{=} U_k \sim U(0,1) \). Combined with the fact that the \( kth \) order statistic of the
standard uniform distribution follows a beta distribution, we have that \( F(q_{jk}) \overset{d}{=} U_{jk} \sim \beta(k, N_j + 1 - k) \)
, where \( \beta(x, y) \) is a beta distribution with parameters \( x \) and \( y \). This yields the following:

\[
E(q_{jk}) = E(F^{-1}(F(q_{jk}))) = E(F^{-1}(U_{jk}))
\]  

(4)

The quantile function for the uniform distribution is the inverse of the CDF. In general terms, the function for quantile \( p \) is written as:

\[
F^{-1}(p) = a - p(b - a)
\]  

(5)

Substituting (5) into (4) yields:

\[
E(q_{jk}) = E(a_j + U_{jk}(b_j - a_j)) = a_j + (b_j - a_j)E(U_{jk})
\]  

(6)
The expected value of $\beta(x, y) = \frac{x}{x + y}$. Substituting this expression along with equation (6) into equation (3) gives us:

$$S_{j,k} = \sum_{k=1}^{R_j} E(q_{j,k}) = \sum_{k=1}^{R_j} (a_j + (b_j - a_j)(\frac{k}{N_j + 1})) = R_j a_j + (\frac{b_j - a_j}{2(N_j + 1)})(R_j(R_j + 1)). \quad (7)$$

Thus, the maximization problem in equation (2) can be written as:

$$\text{MAX}_{R_A, R_B} E(Q) = (\mu_A R_A - R_A a_A - (\frac{b_A - a_A}{2(N_A + 1)})R_A(R_A + 1)) + (\mu_B R_B - R_B a_B - (\frac{b_B - a_B}{2(N_B + 1)})R_B(R_B + 1)) \quad (8)$$

Imposing the constraint $R_A + R_B = \overline{R}$, the Langrangian yields the following first-order conditions:

$$\begin{align*}
\mu_A - a_A - (\frac{b_A - a_A}{2(N_A + 1)})(2R_A + 1) - \lambda = 0  \\
\mu_B - a_B - (\frac{b_B - a_B}{2(N_B + 1)})(2R_B + 1) - \lambda = 0  \\
\overline{R} - R_A - R_B = 0  
\end{align*} \quad (9)$$

To see that the proportional policy satisfies the first-order conditions, note that $N_A = N_B$ (equal-sized sectors), $b_A - a_A = b_B - a_B$ and $\mu_A - a_A = \mu_B - a_B$ (the latter two equalities are the result of our assumption that the variance of teacher quality is equal across sectors – we examine the variance issue below in detail with numerical simulations). Given these equalities, the first-order conditions are clearly met when $R_A = R_B = \frac{\overline{R}}{2}$, which defines the proportional policy. Under the global policy, where in this case $R_A < R_B$ (because $\mu_A > \mu_B$), the first order conditions are not satisfied.

Finally, the second order conditions are defined by the bordered Hessian matrix:
\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & -\frac{b_A-a_A}{(N_A+1)} & 1 \\
1 & 0 & -\frac{b_B-a_B}{(N_B+1)}
\end{pmatrix}
\]  
(10)

where \( |H| = \frac{b_A-a_A}{N_A+1} + \frac{b_B-a_B}{N_B+1} > 0 \), indicating that the proportional policy satisfies the conditions for the maximum.

Of course, the model presented in this section is simplified in a number of ways. Below we use numerical simulations to illustrate the efficiency implications of proportionality under a variety of more complicated circumstances, including cases where incumbent teacher quality is not known and must be estimated (which introduces imprecision and potentially bias), where teachers are drawn from an alternative distribution (normal), and where the variance of teacher quality differs across sectors (parameterized based on available empirical evidence). Our numerical simulations also allow us to examine a variety of other issues including turnover costs and policy persistence. A consistent finding in the analysis that follows is that the proportional policy is more efficient – or at least no less efficient – than the global alternative under a variety of plausible evaluation conditions.

3.2 Incumbent and Replacement Teacher Quality

The efficiency of the proportional policy illustrated in the previous section depends on the gap in teacher quality across school types for incumbents carrying over for replacement teachers as well. Is it reasonable to expect replacement teachers to exhibit the same quality gaps as incumbents? Below we list four potential mechanisms that might explain documented gaps in observed teacher quality between high- and low-poverty schools:

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).
2. The gaps reflect differences in the quality of leadership across school types (Koedel et al., 2012). This explanation requires leadership quality to influence measured
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., forthcoming; Raudenbush 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.

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 imply a direct connection between current teachers and their replacements (or how their performance is differentially measured). The applicant-pool explanation lends itself most directly to the way that we have framed the problem above, but the “at work” mechanisms, like leadership quality and access to instructional strategies, would also result in a link between observed- and replacement-teacher quality within school types. 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 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.\(^\text{10}\) Thus, our findings will apply regardless of the mechanisms that drive existing quality gaps between teachers in high- and low-poverty schools.

\(^{10}\) This is true in the absence of a policy designed to address the fundamental source of the observed quality gap – in the case where the 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.
4. Simulations and Empirical Analysis

4.1 Generating Simulated Data

Building on work by Winters and Cowen (2013), we generate simulated data to extend our investigation of the efficiency implications of proportionality and bring in various aspects of the real-world evaluation problem. We begin by defining each teacher's annual estimated performance as the sum of three components:

\[ \hat{y}_j = q_j + \delta_j + \eta_j \]  

(11)

In equation (11) \( q_j \) is a time-invariant performance measure, \( \delta_j \) varies from year to year and is independent across years (\( \delta_j \) can be viewed as reflecting year-to-year teacher-classroom match effects and/or natural variation in teacher performance over time), and \( \eta_j \) 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 (2013) is that they draw their analog to \( \eta_j \) 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 \( \eta_j \) truly reflects sampling variance.

We randomly assign students to teachers in our simulations.\(^{11}\) The test score for student \( i \) with teacher \( j \) in year \( t \) can be written as:

\[ Y_{it} = \alpha_i + (q_j + \delta_j) \]  

(12)

Equation (12) shows that student scores are a function of a student-year specific component, \( \alpha_i \), and teacher assignments. \( \alpha_i \) captures a number of factors that influence student test scores, most notably

---

\(^{11}\) We assign students to teachers randomly to maintain focus on our research question with limited distractions, as in Schochet and Chiang (2013). Winters and Cowen (2013) also effectively randomly assign students to teachers. 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 (2014a), and Kane et al. (2013) offers reason for optimism about available models, but a detailed discussion of this issue is beyond the scope of this paper.
student ability and test measurement error (Boyd et al., 2013). For our purposes, $\alpha_u$ is best viewed as the residual variance in student test scores after accounting for the role of teachers. Note that equations (11) and (12) are linked through $\alpha_u$ because $\eta_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \alpha_{ui}$, where $N_j$ 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 (11) – we use one of their parameterizations where $\sigma_q = 0.15$, $\sigma_\delta = 0.15$, $\sigma_\eta = 0.21$. Implicit in their setup is that the variance of student scores is normalized to one, and we use the same normalization (thus $\sigma_\eta = 1.0$). With random assignment of students to teachers the expected variance of $\eta$ is driven by the student/teacher ratio. The above-specified variance of $\eta$ is achieved when we set this ratio to 20, which along with the above-specified variances of $q_j$ and $\delta_j$ results in a year-to-year correlation in $\hat{\gamma}_j$ of 0.25. This parameterization uses a plausible value of $\sigma_q$ and also produces plausible estimates of $\sigma_\eta$ (e.g., see Hanushek and Rivkin, 2010). The findings we present below 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.\(^{12}\) We again group teachers into two school types: type-A and type-B schools. We generate data for 12,000 students and 600 teachers and set the student-teacher ratio to 20 for all 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

\(^{12}\) Winters and Cowen (2013) briefly consider a scenario with differentiated schools; however, they do not consider the proportionality issue in their study.
allow school effects to enter into the data generating process directly in any way beyond the distinction between type-A and type-B schools.\textsuperscript{13}

4.2 Baseline Simulation Results

In this section we provide the numeric analog to Section 3.1 using the simulated data. We begin by drawing teachers in type-A and type-B schools from normal quality distributions defined by $q^A_j \sim N(0.05, (0.15)^2)$ and $q^B_j \sim N(-0.05, (0.15)^2)$, respectively. Thus, the expected gap in quality between low- and high-poverty schools is 0.10 standard deviations of the student achievement distribution, parameterized in the first moment. This is perhaps an implausibly large difference in quality across school types, but it is useful for illustration. We consider more moderate quality gaps later on.\textsuperscript{14}

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

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

\textsuperscript{13} 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, $\alpha \sim N(0, \sigma^2_\alpha)$. Our simulation framework is flexible enough to incorporate heterogeneous school effects and student sorting. However, only sorting and heterogeneity across school types is directly relevant for our analysis, and per Section 3.2, differences by school type are captured in our framework by what we refer to for ease of presentation as differences in teacher quality. Put differently, if one of the other mechanisms discussed in Section 3.2 is present, we could simply re-label the teacher-quality gaps.

\textsuperscript{14} Imposing the quality gap across school types also increases $\sigma_q$ as measured across all schools. In principle we could reduce $\sigma_q$ 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).
student test scores under the global policy and 0.0263 under the proportional policy. The achievement gains are concentrated among students taught by the replacement teachers, which means that the average gain in $q$ per removed teacher is equal to ten times the system-wide achievement gain: 0.250 and 0.263 for the global and proportional policies, respectively.

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, for both the global and proportional policies, the figure shows (1) the average $q$-value for the marginally-removed teacher in each school-type (striped bars), and (2) the expected gain in achievement owing to the marginal replacement (solid bars). Using the terminology from section 3.1, these are the values of $r^m_j$ and $\Delta r^m_j = (\mu_j - r^m_j)$ for $j = A,B$ under each policy.

The first set of bars in the figure show that the global policy is structured to equate the values of $r^m_A$ and $r^m_B$. But because of the quality difference of replacements, this necessarily results in $\Delta r^m_A > \Delta r^m_B$, implying that total achievement can be increased by transferring the marginal removal from a type-B to type-A school. In contrast, under the proportional policy $r^m_A < r^m_B$, which facilitates the equating of $\Delta r^m_A$ and $\Delta r^m_B$, and thus efficient removals.

The efficiency gain from proportionality is driven by shifting marginal removals out of type-B schools and into type-A schools. Based on the parameterization of the simulations thus far, the global policy removes 15.4 percent of teachers in type-B schools and 4.6 percent of teachers in type-A schools, whereas by construction the proportional policy removes 10 percent of teachers from each school type. Table 1 shows that while the net effect of shifting removals between school types is to increase achievement overall, the achievement gain in type-B schools is lower with the proportional policy relative to the global policy. 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, 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 are 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.\(^\text{15}\)

Below we elaborate further on the equity issue, which is complex. For example, countering the achievement results for type-B schools presented thus far, Ehlert et al. (forthcoming) discuss the benefits of proportionality in terms of mitigating the potentially adverse effects of global evaluations on recruitment into high-poverty schools in the first place. In addition, an important aspect of the real-world problem yet to be incorporated into our simulations is that turnover is costly, and moreover, turnover costs are asymmetric between high- and low-poverty schools (Ronfeldt et al., 2011). We return to these and other equity issues in Section 5.8.

5. Extensions

In this section we present results from a number of extensions of our baseline simulation framework. The extensions are designed to incorporate various dimensions of the real-world evaluation problem. Table 2 shows results for each extension in the same format, reporting results analogous to what we show for the “Achievement Gain” and “Combined Achievement Gain (weighted)” in Table 1. The results from Table 1 are replicated in the first row of Table 2 for ease of comparison. A common theme in all of the extensions is that the proportional policy is either more efficient, or not meaningfully less efficient, than the global alternative.

\(^{15}\) 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 at 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 50th percentile in the type-B distribution. These teachers are still below the 50th 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.
5.1 Policies Based on Unbiased Estimated $q$

We begin with a case where $q$ is not known and 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 only because of statistical imprecision. In real-world applications, this scenario is informative if we believe that the model from which we estimate $\hat{q}$ is sufficiently rich so that there is no bias in the teacher-quality estimates.

Row 2 of Table 2 presents results when we remove teachers based on $\hat{q}$ rather than $q$, with removals depending on single-year estimates of $\gamma_{jt}$ for each teacher. As anticipated, the noise in the performance signals attenuates the achievement effects for both policies relative to the case where removals are based on known $q$ (row 1). The average system-wide gains in achievement under the global and proportional policies are 0.0127 and 0.0129, respectively.\textsuperscript{16} The efficiency gain from using the proportional policy remains despite being reduced by statistical imprecision. Returning to the marginal-removal logic from above (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 removals are effectively random.

5.2 Allowing for Biased Estimates of Teacher Performance

In the previous section we introduced noise into the estimation process, but not bias. 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

\textsuperscript{16} 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.
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.\(^\text{17}\)

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. (forthcoming) 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 favor advantaged schools and teachers teaching in these schools. The reason is that there may not be sufficient variation to properly identify the coefficients on the variables that control for schooling context, leading to attenuation in these 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 estimates 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).\(^\text{18}\)

We introduce bias into the estimated teacher effects by imposing ad hoc bias terms of +0.02 and -0.02 for teachers in type-A and type-B schools, respectively. Row 3 of Table 2 shows results

---

\(^{17}\) Of course, if bias is an important concern then 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, 2014a; Kane et al., 2013), 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.

\(^{18}\) 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. However, as a practical matter this is not currently how SGPs are used.
analogous to those from the previous section, 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 offers 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. Specifically, the per-student gain in student achievement when removals depend on the global policy is 0.0125, while under the proportional policy the gain is 0.0130.

5.3 Allowing for Natural Attrition

Thus far we have not allowed for teacher attrition other than attrition driven by the removal policies. In reality attrition for other reasons is high. In this section we layer the removal policies on top of natural teacher attrition. This will dull the effects of the policies because some teachers who are targeted for removal will leave on their own. We allow natural attrition to depend 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, with the most and least effective teachers being the most likely to leave.

Row 4 of Table 2 shows how incorporating natural teacher attrition into the analysis affects our findings. We use estimated teacher quality to make the removal decisions in row 4, without bias, which means that the results in row 2 of Table 2 serve as the baseline comparison case without natural attrition. When we allow for natural attrition the total achievement effects of both policies are smaller because fewer teachers are removed involuntarily, but the most efficient policy continues to be the proportional policy.

5.4 Using Multiple Years of Data to Inform Policy Action

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 row 5 of
Table 2 we incorporate this dimension of real-world evaluations by using estimates of teacher effectiveness based on three years of data to determine removals. 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 with three years of data are ranked and the bottom 10 percent are removed. Teachers who replaced natural exiters at the end of years one and two are excluded from the removal program. The effect of using three years of data as opposed to a single year can be seen by comparing the results in row 5 to the results in row 4, which are similar except for the number of years of data used to measure teacher quality. Unsurprisingly, the performance of both the global and proportional policies is improved using the three-year estimates. This is because the roles of $\delta_{jt}$ and $\eta_{jt}$ in determining the removal decisions are reduced. The proportional policy remains the most efficient, albeit by a small margin.$^{19}$

5.5 Changing the Gap in $q$ Across School Types

In Row 6 of Table 2 we lower the gap in average teacher quality across school types to 0.05, maintaining the other evaluation conditions from row 5. The smaller quality gap is closer in magnitude to the gaps reported by Sass et al. (2012) between high- and low-poverty schools in Florida and North Carolina.$^{20}$ The efficiency gain from the proportional policy remains small in row 6, and is identical up to the fourth decimal place to what we estimate with the larger gap. Returning to the logic from Sections 3.1 and 4.2, 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.

$^{19}$ Note that there are fewer total removals in row 5 relative to row 4 because the removal policies are implemented conditional on teachers who have three years of data. This reduces the achievement effect for both policies, but the loss is more than offset by the gain that comes from the improved precision in estimated teacher quality.

$^{20}$ The smaller gap is also consistent with gaps in teacher quality between free/reduced-price lunch eligible and non-eligible 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.
5.6 *Policy Permanency*

The above results are all based on “single shot” removal policies. The efficiency gain from the proportional policy will be amplified if it is iterated over time (similarly to Winters and Cowen, 2013). We examine policies that iterate for five years in row 7. For the iterative policies we mimic the general evaluation conditions in row 4 (i.e., we set the quality gap to 0.10, estimate $q$ based on one year of data, and we allow for natural teacher attrition). Thus, the increased gains from iterating the policies can be seen by comparing the results in row 7 to the results in row 4.

The iterative global and proportional policies have much larger achievement effects than their “single shot” counterparts. Specifically, the achievement gain under the global policy jumps from an average gain per student of 0.0095 (row 4) standard deviations of student test scores for a single year to 0.0343 (row 7) standard deviations after iterating for five years; under the proportional policy the gain jumps from 0.0098 to 0.0350. The efficiency gain from proportionality is larger when the policies iterate.

5.7 *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. 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).\(^\text{21}\) Nonetheless, we take the Sass et al. estimates at face value and calibrate the data generating process

\(^{21}\) 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.
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 10th, 25th, 50th, 75th and 90th 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 10th percentile), while upper tail teachers in low-poverty schools are actually less effective than upper tail teachers in high-poverty schools (the gap is approximately -0.021 at the 90th percentile). Sass et al. also estimate that there is more 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).

Rows 8 and 9 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.22

Unlike in the preceding analysis, it is not certain that the proportional policy will be the most efficient in rows 8 and 9. The efficiency rationale illustrated in Sections 3.1 and 4.2 depends on the

---

22 We use the 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 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 distributional differences across school types.
variance of teacher quality being the same in both school types. In rows 8 and 9, while the marginal-removal intuition from above continues to work in favor of proportionality, the higher variance in type-B schools works against it. This is because the higher variance 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 (row 8), the proportional policy is slightly less efficient overall. Using the North Carolina calibration, the efficiency gain from proportionality is effectively zero.

These results show 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 calibrations using 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.

5.8 Equity

As noted in Section 4.2, the constraint on the number of removals is such that the proportional policy lowers the achievement gain in type-B schools by shifting removals across sectors. Again, the reason is that 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. This feature of proportionality is reflected through the first nine rows of Table 2, which consistently show that while the total achievement gain is higher under the proportional policy, the achievement gain for type-B schools is higher under the global policy.

The equity issue in this context is complicated. While the proportional removal policies that we have considered thus far result in lower achievement at type-B schools than their global
counterparts, other considerations push in the opposite direction. For example, the efficiency gain from a symmetric policy that awards retention bonuses to highly-effective teachers will be accompanied by an improvement in performance at type-B schools (at the expense of type-A schools). To see this, consider a retention bonus program targeted at the top 10 percent of teachers. Based on global teacher rankings and with perfect information as in Figure 1, teachers in type-A schools will be overrepresented among bonus recipients to the same degree that they are underrepresented in the global removal policy. 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 from above. In addition, the absolute number of retained effective teachers in type-B schools increases, which increases achievement at these schools.23

A second 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 comparisons are proportional and performance matters, teachers will be encouraged at the margin to shift to type-B schools if they are more likely to be rewarded in those schools. Ehler et al. (forthcoming) also make this point.

Returning to context of our 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-based achievement penalty will partly offset any adverse equity effects. Furthermore, evidence from Ronfeldt et al. (2011) indicates that turnover is more costly

---

23 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 we are not aware of research evidence that provides this information. The issue is that a retention bonus will not ensure retention in the same way that a removal policy will (mostly) ensure removal.
at high-poverty schools, implying an additional mechanism by which the proportional removal policy will improve efficiency overall.

We incorporate turnover costs into our analysis in row 10 of Table 2. We report results from a single-year removal policy where turnover costs at type-A and type-B schools are parameterized based on Ronfeldt et al. (2011).24 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 results without turnover costs are shown in row 4 (turnover costs are applied to turnovers caused by both natural and policy-based teacher attrition, although this is not a qualitatively important detail for comparing the proportional and global policies).

Of course, incorporating turnover costs lowers the gains from the removal policies because both policies increase turnover. Still, even with the turnover penalties in place, row 10 shows that the policies meaningfully improve student achievement. Moreover, moving from the global to proportional policy, 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 row 10 of the table, but unlike in the previous results, the total efficiency gain does not come at the expense of type-B schools.

Finally, row 11 expands on the scenario in row 10 by allowing the policies to iterate for five years. Analogous results without the turnover penalty are shown in row 7. Row 11 is more realistic than row 10 because it incorporates the fact that 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 row 11 show that both policies improve achievement, and that the proportional policy is again the

24 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.
most efficient. The equity tradeoff re-emerges, but the loss for type-B schools in moving to the proportional policy is decreased by the reduced turnover costs.

6. Additional Considerations

6.1 Non-Test-Based Measures

Our efficiency findings will apply to any performance-based measure ranging from classroom observations to student surveys to value-added. The reason is that all of the mechanisms that can explain systematic differences in teacher ratings by these metrics across schooling contexts imply that the differences will persist for replacement teachers as well (per Section 3.2). The logic from our test-based simulations translates directly to other measures of teacher performance.\(^{25}\)

6.2 Teacher Churning

As noted in Section 5.8, an equity-enhancing feature of proportional evaluations is that they may improve the prospects for high-poverty schools in the labor market. All else equal, high poverty schools will be more appealing to teachers under the proportional policy because teachers will be able to achieve a higher rating. Unlike most aspects of the teacher labor market, which favor advantaged schools, a proportional evaluation policy favors disadvantaged schools.

However, one can also imagine proportionality-induced teacher churning that would harm equity or be unproductive in other ways. For example, consider the case where there are two identical, neighboring districts where one district adopts a proportional policy and the other adopts a global policy. Rather than encouraging teachers in low-poverty schools to move to high-poverty schools within the proportional-policy district, the proportional policy may encourage teachers in low-poverty schools to change to the other district. The general equilibrium effects are complicated, and depend on the elasticity of teacher labor supply across several dimensions.

\(^{25}\) Teacher performance measures need not derive from models of student achievement to be constructed as proportional. Proportionality can always be achieved via \textit{ex post} regression adjustment at the teacher level. For a recent, related discussion in the context of classroom observations see Whitehurst, Chingos and Lindquist (2014).
As a second example, consider a large school district that contains high- and low-poverty schools. One aspect of the proportional policy is that it will generate additional vacancies at low-poverty schools (relative to the global policy). For a district governed by a single labor contract this may be problematic. Specifically, if the additional vacancies at low-poverty schools are desirable to teachers, and experienced teachers in high-poverty schools are priority candidates for these positions, teacher churning within the system could completely undo the benefits of proportionality and may have adverse equity implications.

In short, there are a number of dynamic labor-market issues that would merit consideration if policymakers were to implement a proportional policy. One way to guard against some types of inefficient churning is to reduce labor-contract rigidities in teaching. Another is to implement evaluation policies more broadly – for example, at the state rather than district level – to lessen concerns about strategic cross-district teacher mobility.

7. 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, combined with the fact that access to effective teachers meaningfully affects students’ immediate and longer-term outcomes (Hanushek and Rivkin, 2010; Chetty, Friedman and Rockoff, 2014b). 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 evaluations. Although we cannot hope to capture all aspects of the teaching

26 There are several recent studies on the general issue of teaching-contract rigidities – two recent studies on “last in, first out” policies are Boyd et al. (2011) and Goldhaber and Theobald (2013). Reducing labor-contract rigidities for educators was a focal point of the recent Vergara v. California court case.
profession and the associated dynamics of the labor market with our simulations, 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. The efficiency gains that we document under real-world evaluation conditions are small, but 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 et al., forthcoming), our findings point to proportional evaluations as being a useful alternative for educational administrators charged with developing and implementing teacher evaluation systems.

References


Figure 1. The Quality of the Marginally Removed Teacher, and the Expected Gain in Quality from the Marginal Removal and Replacement, at Each School Type Using the Global and Proportional Policies. Reported Values are Averaged across 1,000 Simulations.

Notes: As described in the text, the expected quality for replacement teachers in the figure is 0.05 at type-A schools and -0.05 at type-B schools. The standard deviation of quality within each sector is 0.15. The figure shows that the proportional policy equates the gain from the marginal removal across sectors (solid bars). Table 1 shows the total gains in workforce quality under the global and proportional policies using the scenario illustrated in this figure.
Table 1. Gains in Workforce Quality and Student Achievement after Removing 10 Percent of the Teaching Workforce based on Persistent Effectiveness ($q$), Perfectly Observed.

<table>
<thead>
<tr>
<th></th>
<th>Removal Policy based on Global Rankings</th>
<th>Removal Policy based on Proportional Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type-A Schools</td>
<td>Type-B Schools</td>
</tr>
<tr>
<td>Average Quality in the Absence of Policy</td>
<td>0.0500</td>
<td>-0.0499</td>
</tr>
<tr>
<td>Average Quality with Policy</td>
<td>0.0646</td>
<td>-0.0144</td>
</tr>
<tr>
<td>Achievement Gain</td>
<td>0.0146</td>
<td>0.0355</td>
</tr>
<tr>
<td>Combined Achievement Gain (weighted)</td>
<td>0.0250</td>
<td>0.0263</td>
</tr>
<tr>
<td>Number of Teachers Removed</td>
<td>13.894</td>
<td>46.106</td>
</tr>
<tr>
<td>Average Gain in $q$ per Replacement</td>
<td>0.315</td>
<td>0.231</td>
</tr>
<tr>
<td>Average Gain in $q$ per Replacement (weighted)</td>
<td>0.250</td>
<td>0.263</td>
</tr>
</tbody>
</table>

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. Average Student Achievement Gains for the Global and Proportional Removal Policies under Different Evaluation Conditions, by School Type and Combined.

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Global Rankings</th>
<th>Proportional Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type-A (combined gain)</td>
<td>Type-B (combined gain)</td>
</tr>
<tr>
<td>1. Known $q$ (replication of results reported in Table 1)</td>
<td>0.0146 (0.0250)</td>
<td>0.0355 (0.0263)</td>
</tr>
<tr>
<td>2. Unbiased Estimates of $q$ Based on 1 Year of Data</td>
<td>0.0100 (0.0127)</td>
<td>0.0153 (0.0129)</td>
</tr>
<tr>
<td>3. Biased Estimates of $q$ Based on 1 Year of Data</td>
<td>0.0089 (0.0125)</td>
<td>0.0162 (0.0130)</td>
</tr>
<tr>
<td>4. Unbiased Estimates of $q$ Based on 1 Year of Data, Natural Teacher Attrition</td>
<td>0.0077 (0.0095)</td>
<td>0.0113 (0.0098)</td>
</tr>
<tr>
<td>5. Unbiased Estimates of $q$ Based on 3 Years of Data, Natural Teacher Attrition</td>
<td>0.0081 (0.0114)</td>
<td>0.0146 (0.0115)</td>
</tr>
<tr>
<td>6. Unbiased Estimates of $q$ Based on 3 Years of Data, Natural Teacher Attrition, Quality Gap Reduced from 0.10 to 0.05</td>
<td>0.0100 (0.0114)</td>
<td>0.0128 (0.0114)</td>
</tr>
<tr>
<td>7. Unbiased Estimates of $q$ Based on 1 Year of Data, Natural Teacher Attrition, Policy Iterates for 5 years</td>
<td>0.0302 (0.0343)</td>
<td>0.0383 (0.0350)</td>
</tr>
<tr>
<td>8. Unbiased Estimates of $q$ Based on 1 Year of Data, No Natural Teacher Attrition, Distributional Gaps in Teacher Quality Across Schools based on Florida</td>
<td>0.0083 (0.0112)</td>
<td>0.0141 (0.0111)</td>
</tr>
<tr>
<td>9. Unbiased Estimates of $q$ Based on 1 Year of Data, No Natural Teacher Attrition, Distributional Gaps in Teacher Quality Across Schools based on North Carolina</td>
<td>0.0109 (0.0124)</td>
<td>0.0138 (0.0124)</td>
</tr>
<tr>
<td>10. Unbiased Estimates of $q$ Based on 1 Year of Data, Natural Teacher Attrition, Turnover Penalty Imposed</td>
<td>0.0058 (0.0049)</td>
<td>0.0040 (0.0052)</td>
</tr>
<tr>
<td>11. Unbiased Estimates of $q$ Based on 1 Year of Data, Natural Teacher Attrition, Turnover Penalty Imposed, Policy Iterates for 5 years</td>
<td>0.0258 (0.0280)</td>
<td>0.0301 (0.0290)</td>
</tr>
</tbody>
</table>

Note: This table reports results analogous to those shown in the rows labeled “Achievement Gain” and “Combined Achievement Gain (weighted)” in Table 1 for a variety of evaluation conditions. Within each cell, the top two numbers are the school-type-specific achievement gains, and the bottom number (in parenthesis) is the combined achievement gain across all schools. Unlike in Table 1, the achievement gains are not equal to 10 percent of the average gain in $q$ per removed teacher for all scenarios shown in this table, although for some scenarios this remains the case. More information is available from the authors upon request.