

Proportionality in District, School and Teacher Evaluation Systems

Cory Koedel

Summary

- Rigorous district, school and teacher evaluations are increasingly common across the United States, but there is still considerable controversy over how best to design evaluation systems.
- Proportionality is a design feature of educational evaluations. Proportional evaluations fully level the playing field between districts, schools and teachers in different educational environments. As a result of the equally-circumstanced comparisons, rankings are proportional throughout – for example, teachers working in advantaged and disadvantaged schools are equally likely to be identified as top and bottom performers in a proportional evaluation system.
- Critics of the recently-defeated Amendment 3 in Missouri, which would have linked teacher evaluations to student performance, argued against the amendment in part because of concerns over “poor results for educators of disadvantaged students who might score poorly regardless of instruction” (Ballentine, 2014). Proportional evaluations entirely mitigate this concern. More generally, proportionality is a desirable property of educational evaluations given key policy objectives of emerging evaluation systems.

Introduction

State and local education agencies across the United States are working to design and implement more-rigorous evaluation systems for districts, schools and teachers based on student outcomes, sometimes with high stakes attached. A motivation for implementing these new systems is the consistent finding in research that schools and teachers differ dramatically in quality (Betts, 1995; Chetty, Friedman and Rockoff, 2014; Hanushek and Rivkin, 2010). In many locales, the large quality differences that we know exist between schools and teachers are not reflected in historically available evaluation metrics (e.g., in the case of teachers, see Weisberg et al., 2009), and researchers have had great difficulty linking performance differences to readily-observable characteristics (Betts 1995; Nye, Konstantopoulos and Hedges, 2004). The appeal of newly-emerging evaluation systems is their ability to better differentiate districts, schools and teachers based on performance.

Although there is a growing consensus among researchers and policymakers that students in K-12 schools stand to gain substantially from the implementation of more-rigorous, performance-based evaluation systems in public education, there are ongoing policy debates regarding how best to design these systems. This policy brief introduces and discusses a design feature referred to as proportionality. Proportional evaluation systems force comparisons to be between equally-circumstanced units (e.g., schools or teachers). As a result of these forced comparisons, rankings are proportional to the evaluation sample throughout. For example, teachers working in low-poverty and high-poverty schools are equally likely to be identified as top and bottom performers in a proportional teacher evaluation system by construction.



Cory Koedel is Associate Professor of Economics and Public Policy at the Harry S Truman School of Public Affairs, University of Missouri.

This is in contrast to “global” evaluation systems, which can generate “disproportional” rankings – for example, teachers working in low-poverty schools might be over-represented in the top quartile of rankings and teachers working in high-poverty schools over-represented in the bottom quartile.

Figure 1 provides a visualization of the difference between global and proportional rankings using school-level data from Missouri elementary and middle schools. The graphs in the figure are taken from Ehlert et al. (forthcoming). They plot “global” and “proportional” performance measures for schools against their shares of students eligible for free/reduced-price lunch (a proxy for student disadvantage). A regression line that characterizes the overall relationship is shown in each graph.

The global measures are estimated using what Ehlert et al. (forthcoming) refers to as a “one-step” value-added model and the proportional measures are estimated using what they refer to as a “two-step” value-added model. The value-added models are estimated using student achievement data from the Missouri Assessment Program (MAP) test from 2007-2011. The two-step model enforces proportionality by statistically controlling for observed student and school characteristics prior to estimating the performance measures for schools. The interested reader can refer to Ehlert et al. (forthcoming) for technical details about the statistical models.

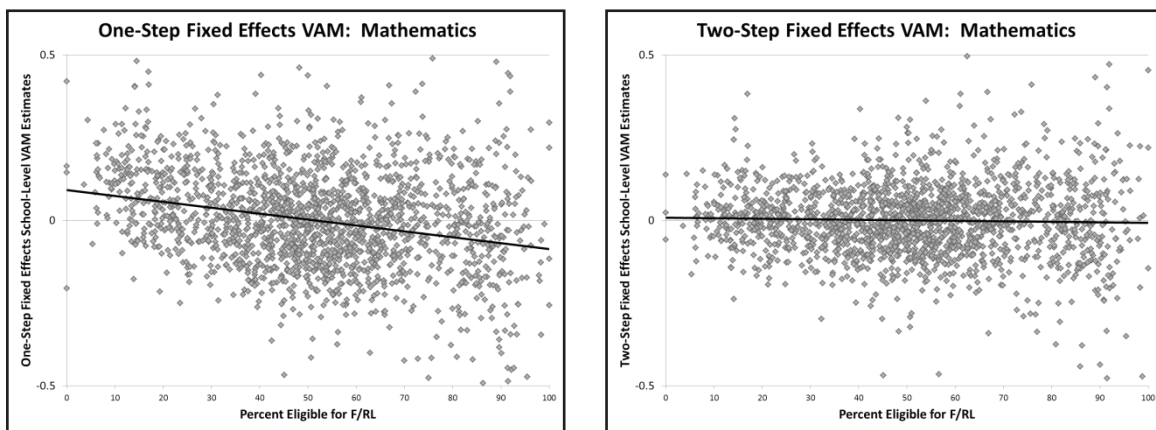


Figure 1. Both graphs compare estimates of school performance to the share of students eligible for free/reduced-price lunch (a proxy for student disadvantage). The panel on the left shows estimates from a model designed to produce global rankings (under some assumptions); the panel on the right shows estimates from a model designed to produce proportional rankings

The flat line in the proportional-rankings graph, again based on output from the statistical model discussed by Ehlert et al., illustrates the effect of imposing proportionality, which is to force equally-circumstanced comparisons. The implication of the flat line is that school rankings are not systematically predicted by the share of students at the school who are eligible for free/reduced-price lunch. This is contrast to the global-rankings graph, where the negatively-sloping line reflects the fact that high-poverty schools are ranked systematically lower than their low-poverty counterparts.

The Benefits of Proportionality in District, School and Teacher Evaluation Systems

A discussion of whether proportional evaluations are desirable requires specifying policy objectives. Ehlert et al. (forthcoming, 2014) specifies three objectives of any educational evaluation system: (1) improve system-wide instruction by providing useful performance signals to teachers and administrators, (2) elicit optimal effort from teachers and administrators, and (3) avoid exacerbating pre-existing inequities in the labor markets faced by high- and low-poverty schools.

Improve the informational value of system ratings. To illustrate the value of proportionality in achieving this objective, consider an evaluation system for schools in a world in which there are just two types of schools, high-poverty and low-poverty, and where low-poverty schools have higher performance on average (based on global ratings as shown in Figure 1). The equally-circumstanced feature of proportional evaluations offers two benefits in this situation. First, even if it is more difficult to improve student outcomes in high-poverty schools relative to low-poverty schools, some high-poverty schools will still perform much better than others, and proportional evaluations will preserve this information. In contrast, global ratings that systematically favor low-poverty schools will make all high-poverty schools look relatively worse, which can mask the fact that some high-poverty schools are performing quite well. A potentially harmful consequence of a system that fails to distinguish high-performing, high-poverty schools is that it may result in a perpetuating cycle of the destruction and re-invention of instructional practices at these schools, whether the practices are effective or not. Put differently, if high-poverty schools that are performing well do not receive a proper signal to indicate as much, they may abandon effective instructional practices. A second benefit of proportionality is that it will better facilitate productive educator-to-educator learning because it makes it straightforward to pair, for example, a low-rated teacher with a high-rated mentor working in a similar educational environment.

Elicit optimal effort from educational personnel. The argument in favor of using proportional evaluations to achieve this objective can be illustrated with a hypothetical scenario where teacher performance ratings will be used to award bonuses to the top 25 percent of teachers (note that many policies surrounding school and teacher evaluations nationally have some competitive element along the lines of this example – for example, see Dee and Wyckoff, 2013). A clear prediction from the economics research literature is that in such a situation, equally-circumstanced comparisons will result in the largest cumulative effort increase across the workforce. Intuitively, the reason is that everyone will feel like they have a chance to win (or lose). In contrast, using global evaluations that systematically favor teachers working in particular environments (e.g., low poverty schools) will produce a weaker effort response because it will place teachers in competition with others against whom they have no hope of winning (or losing). This will weaken effort incentives for everyone (e.g., see Calsamiglia, Franke, and Rey-Biel, 2013; Schotter and Weigelt, 1992).

Avoid worsening pre-existing labor market inequities in public education. These inequities have been well-documented in research (e.g., see Boyd et al., 2005; Jacob, 2007). In short, high-poverty schools are already at a recruiting disadvantage relative to their low-poverty counterparts for a number of reasons. As stakes become attached to educational evaluations, systems that

disproportionately identify teachers in high-poverty schools as “losers” will make positions at these schools even less desirable to prospective educators. Policymakers should proceed cautiously with implementing an evaluation system that will further degrade the pecuniary and non-pecuniary benefits associated with working in challenging educational environments. An important feature of a proportional system is that the “winners” and “losers” from the evaluation will be broadly representative of the system as a whole. High-poverty schools will not be over-represented in the underperforming category.

Based on the goal of satisfying the above-stated policy objectives, proportionality is a desirable design property of educational evaluation systems. Proportionality also offers additional benefits in ways that are more difficult to quantify, but important nonetheless. As one example, Koedel and Li (2014) discuss a situation that arose in Boston where African American teachers were significantly more likely than white teachers to be identified as low-performing (for details see Vaznis, 2013). 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 growth in student achievement (e.g., see Dee, 2004). A proportional model can mitigate differences in teacher assessments that fall along this and other contextual lines.

Concerns

Perhaps the biggest concern that has been raised about imposing proportionality in educational evaluations is that it will hide the potentially inferior performance of districts, schools and teachers in high-poverty areas. However, there is no reason that information needs to be hidden. For example, proportional evaluations can be reported side-by-side with test-score levels. The reporting on test-score levels will allow policymakers to clearly see absolute differences in achievement across schools and proficiency gaps that are unadjusted for student or school characteristics, regardless of how evaluations are performed. Dual reporting of proportional evaluations in conjunction with information about absolute achievement levels is desirable because it allows for the transmission of useful instructional signals. For example, a high-poverty school that is performing well can be encouraged to continue to refine and improve an already-effective instructional strategy in terms of raising test scores compared to similar schools, but still be reminded that students are not scoring sufficiently high relative to an absolute benchmark. The latter information need not disappear in any evaluation framework.

A related concern is that imposing proportionality effectively lowers expectations for socioeconomically-disadvantaged students. However, it is important to recognize that setting expectations for individual students is not the purpose of an evaluation system for educational personnel. The purpose is to achieve key policy objectives like the ones outlined above. Philosophically, policymakers may not want to lower expectations for disadvantaged students. If this is the case, then the proper approach to student-level evaluation is to set fixed success benchmarks for all students and evaluate progress toward those benchmarks. Evaluation system design must be considered within the context of the purpose of performing the evaluations.



Conclusion

Amendment 3 in Missouri, which would have linked teacher evaluations to student performance, was decidedly voted down in Missouri this November. However, the national trend is clearly toward a greater reliance on data in education to help inform evaluations of district, school and teacher performance. Although concerns about fairness to teachers and other educational personnel are obviously important, fairness to students should not be overlooked. For example, recent evidence from Chetty, Friedman and Rockoff (2014) shows the disservice to children in K-12 schools associated with placing them in classrooms with ineffective teachers.

This brief has introduced the concept of proportionality in educational evaluations and discussed its merits in terms of reaching well-defined policy objectives. The distinguishing feature of proportional evaluations is that they force equally-circumstanced comparisons. Districts, schools and teachers in different environments face different tradeoffs and have access to different resources. The factors that determine what constitutes effective practice in different schooling environments are also likely to differ. Pretending that these differences do not exist, and designing evaluation systems that compare districts, schools and teachers regardless of circumstance, is inconsistent with achieving the three policy objectives outlined above, which are to: (1) improve system-wide instruction by providing useful performance signals to teachers and administrators, (2) elicit optimal effort from teachers and administrators, and (3) avoid exacerbating pre-existing inequities in the labor markets between high- and low-poverty schools. Given these three policy objectives, proportionality is a desirable property of educational evaluation systems.

A final point, and a caveat, is that the purpose of this brief has been to provide an overview of the proportionality concept, and not to cover every aspect of the evaluation problem in great depth. The interested reader can refer to Ehlert et al. (forthcoming) for technical information regarding the design of proportional evaluation models and for a comparison to available alternatives, along with an extended discussion of many of the issues touched on in this brief. In addition, Koedel and Li (2014) examine the effect of proportional teacher evaluations on total student achievement. They address concerns over the following situation: proportional evaluations could result in teachers working in low-poverty schools who are ranked below teachers working in high-poverty schools, even though the latter teachers are globally less effective. Koedel and Li (2014) show that under a variety of plausible conditions, even in such a scenario it is still the case that implementing a proportional evaluation system results in larger cumulative gains in student achievement than a global evaluation system.

References

- Ballentine, Summer. 2014. Missouri Voters Reject Teacher Tenure Measure. Associated Press (11.04.2014)
<http://www.sfgate.com/news/crime/article/Budget-teacher-measures-head-to-Missouri-voters-5868031.php> (retrieved November 5, 2014)
- Betts, Julian. 1995. Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth. *Review of Economics and Statistics* 77(2), 231-250.
- 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.
- Calsamiglia, Caterina, Jörg Franke and Pedro Rey-Biel. 2013. The Incentive Effects of Affirmative Action in a Real-Effort Tournament. *Journal of Public Economics* 98, 15-31.
- Chetty, Raj, John N. Friedman and Jonah E. Rockoff 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9), 2633-79.
- 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 (forthcoming). Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter? *Educational Policy*.
- Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2014. Choosing the Right Growth Measure: Methods Should Compare Similar Schools and Teachers. *Education Next* 14(2), 66-71.
- Hanushek, Eric A and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100(2), 267-271.
- Jacob, Brian. 2007. The Challenges of Staffing Urban Schools with Effective Teachers. *Future of Children* 17(1), 129-153.
- Koedel, Cory and Jiaxi Li. 2014. The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce. University of Missouri, Department of Economics Working Paper No. 14-02.
- Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. 2004. How Large are Teacher Effects? *Educational Evaluation and Policy Analysis* 26(3), 237-257.
- 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.
- Vaznis, James. 2013. Union Says Teacher Evaluation Plan has Race Bias. *Boston Globe* (04.23.2013).
- Weisberg, Daniel, Susan Sexton, Jennifer Mulhern and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. New York: The New Teacher Project.

