Economics of Education Review 40 (2014) 76–92

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Teacher heterogeneity, value-added and education policy☆

Scott Condie [a,*], Lars Lefgren [b], David Sims [c]

[a] 136 Faculty Office Building, Provo, UT 84602, United States
[b] 182 Faculty Office Building, Provo, UT 84602, United States
[c] 164 Faculty Office Building, Provo, UT 84602, United States

ARTICLE INFO

Article history:
Received 3 May 2013
Received in revised form 9 September 2013
Accepted 21 November 2013
Available online 8 January 2014

JEL classification:
I2
J4

Keywords:
Value-added
Teacher heterogeneity
Education policy

ABSTRACT

This study examines the theoretical and practical implications of ranking teachers with a one-dimensional value-added metric when teacher effectiveness varies across subjects or student types. We create a theoretical framework which suggests specific tests of the standard teacher input homogeneity assumption. Using North Carolina data we show that value-added fails to empirically meet these tests and document that this leads to a large number of teacher misrankings. Thus, critics of potential value-added teacher personnel policies are correct that such policies will terminate many of the wrong teachers. However, we derive the conditions under which such policies will improve student test scores and find that they will almost certainly be met. We then demonstrate that value-added information can also be used to improve student test scores by matching teachers to students or subjects according to their comparative advantage. These matching gains likely exceed those of a feasible, value-added based firing policy.

1. Introduction

In the summer of 2010, the Los Angeles Times made some news of its own by publishing an online database of value-added test scores, linked to teacher names, for thousands of teachers in the Los Angeles Unified school district. Though this generated millions of page hits for the paper's website it also produced substantial controversy about the proper use of value-added test scores in making education decisions.[1] While a few school districts, most notably the New York City schools, have explored making such teacher scores publicly available, a broader trend in many states has been the increasing use of non-publicized value-added scores in making teacher retention, promotion, and compensation decisions. Indeed, in the fall of 2012, the teachers union in the Chicago Public schools went on strike, partly to reverse the increased use of value-added test scores in teacher evaluation.

The controversy surrounding the use of value-added metrics for evaluating teacher performance largely stems from different weights given by observers to two widely known facts. First, there is substantial, measurable heterogeneity in the ability of teachers to raise student achievement. Second, value-added models depend on several assumptions, some of which have been shown to be empirically invalid in some situations.

One group of observers posits the primacy of the first fact, and emphasizes the promise of value-added to finally create an objective metric for ranking teachers and providing them with incentives. The other group emphasizes the costs of mistaken rankings and generally opposes the whole enterprise. This divergence highlights two important shortcomings of the current literature. First, while economists have produced a number of excellent statistical studies that measure the potential benefits of value-added and the distortions introduced by violating the assumptions of random teacher–student assignments,

☆ We thank Bentley MacLeod and Brian Jacob and seminar participants at Brigham Young University and the University of Michigan for helpful feedback.

* Corresponding author. Tel.: +1 8014225306.
E-mail addresses: ssc@byu.edu (S. Condie), l-lefgren@byu.edu (L. Lefgren), davesims@byu.edu (D. Sims).

[1] The teacher ratings and descriptions can be found at http://www.latimes.com/news/local/teachers-investigation/.

0272-7757/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.econedurev.2013.11.009

they have not yet integrated these with a formal model that shows how policymakers could systematically weigh costs against benefits. Second, the natural focus on statistical parameter identification has led value-added researchers to focus on between teacher heterogeneity and consequently on policies that use value-added for comparing effectiveness between teachers and rewarding success along that dimension.

Conceptually, it is not hard to imagine that one teacher may be more effective than another with one type of students, even as the ranking is reversed with a second type of students. Alternatively, one teacher may be particularly effective at teaching mathematics while another excels at teaching children to read. Such differences across teachers have been empirically documented and likely seem self-evident to parents of school-aged children. However, these concerns are generally assumed away or minimized when considering statistical methods to rank teachers. Hence, potential policy levers such as the non-random assignment of students to teachers actually come to be regarded as a nuisance because they threaten the statistical identification of the teacher quality parameter.

In this paper, we explore the promise and pitfalls of using value-added test scores to improve student learning when the traditional assumption of teacher homogeneity across groups is called into question. To do this, we outline a model of student learning that allows for teacher effectiveness to vary across subjects or student types. We next develop and, using data from 4th and 5th grade students in North Carolina, execute tests of the standard value-added assumptions about the heterogeniety of teacher ability across subjects and types of students. We find that the heterogeneity assumptions of the standard value-added model are rejected. At plausible parameter values, value-added provides a potentially misleading pairwise comparison of teachers between 15 and 25 percent of the time. These misrankings generally arise because value-added indicates one teacher to be superior to another when in fact she is better with only one subject or with one type of students. Furthermore, in the neighborhood of a policy-dictated, value-added cutoff, these measures misrank teachers relative to their social value even in the limit in a majority of cases. Hence, a teacher who is fired may generate more social value than one who is retained. Thus, part of the critique of using value-added measures for personnel policies is undoubtedly correct, such policies will unfairly fire a large number of the wrong teachers.

However, as the primary goal of using value-added personnel policies is to improve student test scores, we also derive the conditions under which this will occur and find them to be almost certainly met. Indeed, simulations with our data confirm that value-added policies would in fact improve social welfare under almost any plausible set of parameter values. Under the parameter values we actually observe, a sample policy of firing teachers in the bottom 10 percent of the observed value-added distribution would increase single-year student achievement by about 0.016 standard deviations in reading and 0.030 standard deviations in math.

Nevertheless, the misrankings of teachers under the standard value-added assumptions suggest that there might be a better way to use value-added information to improve student test scores. Indeed, we prove that assigning teachers to subjects or student types with which they match well can increase student welfare relative to random assignment. Our simulations suggest that teacher specialization according to student ability has the potential to increase reading performance of all students by approximately 0.025 standard deviations. The benefits of assigning teachers to specialize in the subject in which they excel are even larger. This policy would raise math achievement by 0.05 standard deviations and reading achievement by 0.03 standard deviations. These matching improvements are all larger than those realized by replacing the bottom 10 percent of teachers in the value-added distribution. Thus, while it raises concerns about the use of value-added in personnel policies, the heterogeneity of teacher effectiveness across subjects and settings also provides a low-cost potential policy to improve the achievement of all student types.

The remainder of the paper proceeds as described below. Section 2 provides information and highlights the most directly relevant prior research on teacher value-added models. It also describes the data we use for our analysis. Section 3 introduces our model of teacher ability rankings, which allows for teacher ability heterogeneity across student types or subjects, and shows how standard value-added measures emerge as a special case under specific assumptions. Section 4 uses our model to illuminate the conditions under which value-added measurement captures a correct ranking of teacher effectiveness, and then tests those conditions using our data. Section 5 discusses the welfare implications of using value-added to rank teachers when these conditions are not met, and presents policy simulations of the student achievement effects of using value-added for teacher de-selection versus other potential interventions. This section also explores the optimal assignment of teachers to subjects and student types. Section 6 concludes. To aid in the exposition, the proofs of all of the results are found in the appendix.

2. Background and data

2.1. Background

The rise of value-added modeling for teacher evaluation has promised a data-driven method to assess teacher effectiveness. These models leverage extensive longitudinal arrays of matched teacher–student data to produce a numerical measure of a teacher's ability to affect student achievement. Using such models Rivkin, Hanushek, and Kain (2005), Rockoff (2004), and Aaronson, Barrow, and Sander (2007) demonstrate that there are substantial, important differences across teachers in the ability to improve student achievement. This variability of teacher effectiveness has further been confirmed in a large experimental setting by Kane and Staiger (2008). These studies suggest that a one standard increase in teacher value-added ability results in a student test score increase