Economics of Education Review 38 (2014) 9–23

Contents lists available at ScienceDirect

Economics of Education Review

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

Value-added models and the measurement of teacher productivity

Tim R. Sass [a,*], Anastasia Semykina [b], Douglas N. Harris [c]

[a] Department of Economics, Georgia State University, 14 Marietta Street NW, Atlanta, GA 30303, United States
[b] Department of Economics, Florida State University, 113 Collegiate Loop, Tallahassee, FL 32306, United States
[c] Department of Economics, 302 Tilton Hall, Tulane University, New Orleans, LA 70118, United States

ARTICLE INFO

Article history:
Received 18 May 2012
Received in revised form 30 October 2013
Accepted 30 October 2013
Available online 7 November 2013

JEL classification:
I21
J24

Keywords:
Teacher productivity
Value added

ABSTRACT

Research on teacher productivity, as well as recently developed accountability systems for teachers, relies on "value-added" models to estimate the impact of teachers on student performance. We consider six value-added models that encompass most commonly estimated specifications. We test many of the central assumptions required to derive each of the value-added models from an underlying structural cumulative achievement model and reject nearly all of them. While some of the six popular models produce similar estimates, other specifications yield estimates of teacher productivity and other key parameters that are considerably different.

1. Introduction

In the last dozen years the availability of administrative databases that track individual student achievement over time and link students to their teachers has radically altered how research on education is conducted and has brought fundamental changes to the ways in which educational programs and personnel are evaluated. Until the late 1990s, research on the role of teachers in student learning was limited primarily to cross-sectional analyses of student achievement levels or simple two-period studies of student achievement gains using relatively small samples of students and teachers.[1] The advent of statewide longitudinal databases in Texas, North Carolina and Florida, along with the availability of micro-level longitudinal data from large urban school districts, has allowed researchers to track changes in student achievement as students move between teachers and schools over time. This in turn has permitted the use of panel data techniques to account for the influences of prior educational inputs, students and schools when evaluating the contributions of teachers to student achievement.

The availability of student-level panel data is also fundamentally changing school accountability and the measurement of teacher performance. In Tennessee, Dallas, New York City and Washington DC, models of individual student achievement have been used for many years to evaluate individual teacher performance. While the stakes are currently low in most cases, there is growing interest among policymakers to use estimates from student achievement models for high-stakes performance pay, school grades, and other forms of accountability. Chicago, Denver, Houston and Washington, DC have all adopted compensation systems for teachers based on student performance. Further, as a result of the federal *Teacher Incentive Fund* and *Race to the Top* initiatives, many more states and districts plan to implement performance pay systems in the near future. Florida is a particularly interesting case as the state has recently adopted a very

* Corresponding author. Tel.: +1 404 413 0150.
E-mail addresses: tsass@gsu.edu (T.R. Sass), asemykina@fsu.edu (A. Semykina), dharri5@tulane.edu (D.N. Harris).

[1] For reviews of the early literature on teacher quality see Wayne and Youngs (2003), Rice (2003), Wilson and Floden (2003) and Wislon, Floden, and Ferrini-Mundy (2001).

0272-7757/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.econedurev.2013.10.003

aggressive teacher accountability system which relies on these panel data techniques.

Measurement of teacher productivity in both education research and in accountability systems is often based on estimates from panel-data models where the individual teacher effects are interpreted as a teacher's contribution to student achievement or teacher ''value-added.'' The theoretical underpinning for these analyses is the cumulative achievement model developed by Boardman and Murnane (1979), Hanushek (1979), and Todd and Wolpin (2003), where current student achievement is a function of a student's entire history of educational and family inputs. However, varying data constraints have led to a wide variety of empirical specifications being estimated in practice. Each empirical specification makes (typically implicit) assumptions about the parameters of the underlying structural model.

Understanding the assumptions being made and how to test their validity are important both for interpreting the estimates from empirical models and for determining whether the estimates are subject to bias. Model misspecification, including omitted variables, can yield parameter estimates that do not represent the constructs of the underlying structural model and are biased.

Two recent studies evaluate various value-added model specifications based on the criteria of minimizing bias in estimated teacher effects.[2] Kane and Staiger (2008) conduct an experiment in which 78 pairs of teachers were randomly assigned to classrooms in the same grade and school. They then compare pairwise differences in estimated teacher effects from the experimental sample to differences in the pre-experiment value-added estimates of the same teacher pairs. Their analysis included estimates derived from seven value-added specifications with varying controls for student heterogeneity. For the five value-added models that accounted for prior-year student achievement, they could not reject the null that the estimated within-pair differences in teacher productivity were equivalent to the differences under random assignment. Random assignment is a key advantage, but experiments also have limitations. They can generally only be implemented on a small scale and only for individuals or institutions that voluntarily participate. For example, in Kane and Staiger (2008), they could only test whether within-school sorting is an issue and only among pairs of teachers that a principal was comfortable randomly assigning students to. The original experiment was recently replicated across cities with a much larger sample, with the same general results, but the participation rate was once again very low and there were significant problems with non-compliance to the randomization (Kane, McCaffrey, Miller, & Staiger, 2013).

Guarino, Reckase, and Wooldridge (2012) generate simulated data under various student grouping and teacher assignment scenarios and then compare the estimates from alternative achievement model specifications to the known (generated) teacher effects. While no specification is superior under all student/teacher assignment scenarios, a model that estimates current achievement as a function of prior-year achievement and observable student and teacher/school inputs is the most robust. The simulation approach has the advantage of producing known ''true'' teacher effects that can be used to evaluate the estimates from alternative models. The disadvantage, however, is that there is no way to know if the selected data generating processes actually reflect the student–teacher matching mechanisms that occur in real-world data. In particular, the data generating processes they employ relies on a number of simplifying assumptions about the underlying cumulative achievement model.

We take a different approach and test the assumptions required to derive empirical value-added models from a structural model of student achievement. This allows us to determine whether the estimates from value-added models have a structural interpretation.[3] The validity of the assumptions is also important because data generation processes used in simulation work rely on (often implicit) assumptions about the underlying structural model of student achievement. The disadvantage, however, is that failure of the underlying assumptions does not necessarily mean value-added models fail to accurately classify teacher performance for accountability. While we cannot directly test the magnitude of bias in value-added models, we can and do conduct simple hypothesis testing and consider how model specification affects the estimated productivity of teachers. By comparing estimated teacher effects across models of varying flexibility, we can evaluate the magnitude of the change in teacher rankings of specific modeling choices, each with differing data and computational costs. If the results are insensitive to modeling choices then one can be less concerned about imposing false restrictions. But this is not what we find. The results are very sensitive to certain types of assumptions.

We begin our analysis in the next section by considering the general form of cumulative achievement functions and the assumptions which are necessary to generate empirical models that can be practically estimated. In that section we also delineate a series of specification tests that can be used to evaluate the assumptions underlying empirical value-added models. Section 3 discusses our data and in Section 4 we present our results. In the final section we summarize our findings and consider the implications for future research and for the implementation of accountability systems.

2. Value-added models and tests

2.1. A general cumulative model of achievement

In order to clearly delineate the empirical models that have been estimated and the assumptions underlying them, we begin with a general cumulative model of

[2] Another branch of recent literature investigates alternative forms of the cumulative achievement function, emphasizing the impact of historical home and schooling inputs on current achievement. See Todd and Wolpin (2007), Ding and Lehrer (2007), Andrabi, Das, Khwaja, and Zajonc (2011) and Jacob, Lefgren and Sims (2010).

[3] The origins of our work are Harris and Sass (2006), though the analysis has evolved considerably since that original working paper.