ACT Working Paper Series

Will Courts Shape Value-Added Methods for Teacher Evaluation?

Michelle Croft and Richard Buddin

WP-2014-2
April 2014

ACT®

Will Courts Shape Value-Added Methods for Teacher Evaluation?

Abstract: As more states begin to adopt teacher evaluation systems based on value-added measures, legal challenges have been filed both seeing to limit the use of value-added measures (*Cook v. Stewart*) and others seeking to require more robust evaluation systems (*Vergara v. California*). This study reviews existing teacher evaluation systems and examines validity evidence supporting the use value-added scores as part of the teacher evaluation. We discuss key aspects of ongoing teacher evaluation lawsuits in California and Florida and assess issues for evaluating the legality of teacher evaluation systems.

Keywords: teacher value-added, legal issues in education, teacher evaluation

Introduction

Teacher evaluation has traditionally relied on occasional, brief classroom observations by school principals or other administrators. Nearly all teachers received the highest ranking, irrespective of the academic progress of their students (Weisberg et al., 2009). In the past decade, researchers have pioneered so-called value-added assessment methods that delineate teacher effectiveness in terms of their success at improving student academic outcomes in their classroom for one grade to the next (Rivkin, Hanushek, and Kain, 2005; Gordon, Kane, and Staiger, 2006; Harris and Sass, 2006; Aaronson, Barrow, and Sander, 2007; Clotfelter, Ladd, Vigdor, 2007; Koedel and Betts, 2007; Jacob and Lefgren, 2008; Kane and Staiger, 2008; Kane, Rockoff, and Staiger, 2008; Buddin and Zamarro, 2009; McCaffrey, Sass, Lockwood, and Mihaly, 2009; Kane, McCaffrey, Miller, and Staiger, 2013). These studies have found large variation in teacher effectiveness using value-added assessment methods and have suggested that teacher quality could be enhanced by using value-added scores as an element of teacher evaluation. However, the use of teacher value-added measures in personnel decisions remains contentious. Opponents argue that the scores from student achievement tests were not designed to measure teacher performance and the chance of misclassification is too great (Rothstein, et al., 2010). Proponents contend that misclassification is no greater than in other professions and that value-added methods are better able to differentiate among teachers than existing measures (Glazerman, et al., 2010; Goldhaber & Loeb, 2013).

The U.S. Department of Education (U.S. DOE) has sided with the proponents, and through the Race to the Top competition and the ESEA Flexibility Waiver program has prompted states to

use teacher value-added methods as a portion of teacher evaluation systems. According to the ESEA Flexibility timeline, nearly half of states plan to begin using teacher evaluation systems that include teacher value-added methods to inform their personnel decisions (13 states in 2014-2015, 9 states in 2015-2016, and 2 states in 2016-2017) (U.S. Department of Education, 2013). Most recently, Secretary Duncan announced that states may delay implementation until no later than 2016-2017 so that states have more time to implement college readiness standards and incorporate data from implementation of these standards into the value-added assessments. Thus, the majority of states will be using some type of teacher value-added methodology in personnel decisions in the next few years. The efficacy of these reforms will depend on the ability of states and district agencies to design and implement new and complex management systems that incorporate value-added assessments of teachers.

Within this context, the legality of such systems is an important consideration and is already being questioned. Two prominent lawsuits are challenging how teachers should be evaluated. A Florida lawsuit is challenging the legality of recent state legislation that uses student achievement growth as a factor in teacher evaluation (*Cook v. Stewart)*. An additional lawsuit in California is challenging the state's teacher personnel policies (i.e., hiring, dismissal, and layoff decisions) and asking the courts to require individual teacher success at improving student achievement as part of personnel decisions (*Vergara v. California.*). The evidence and outcomes from these lawsuits will provide key benchmarks for other states as they revise teacher evaluation practices over the next several years.

This paper includes three sections. The first provides background on teacher evaluation systems and addresses some of the criticisms of including value-added scores as part of the teacher evaluation. Part two summarizes the two legal challenges focusing on the plaintiffs' arguments and the design of the current evaluation systems. Part three provides design considerations with the emphasis on the factors that the court may consider when evaluating the legality of teacher evaluation systems.

1. Teacher Evaluation Systems and Value-Added Assessment

Current Teacher Evaluation Systems

Teacher evaluations are traditionally based on short classroom observations by school principals or other school administrative personnel. The evaluations are based on instructional practices in the classroom that are believed to be associated with student learning. Data suggests that administrators typically give the highest rating to nearly all teachers, even in classrooms where students are making little or no academic progress (Weisberg et al., 2009). About 75 percent of teachers receive no advice on how to improve instruction, and many administrators have never failed to renew a probationary teacher (Weisberg et al., 2009).

In Los Angeles, for example, the instructional practices of teachers are evaluated annually for new teachers and semiannually for tenured teachers (Newton, 2004). School administrators observe teachers and gauge whether they meet expectations on 25 specific instructional items. Based on these evaluations, fewer than 2 percent of teachers are rated as below standard and targeted for special professional development programs. Newton (2004) reported that over 90 percent of teachers received no negative ratings on any of the 25 items.

These subjective teacher evaluations, like those in most districts, have several limitations. First, the evaluations measure teacher effectiveness against a nominal standard of how teachers should "perform" and not against a measure of how much students actually learn. Ideally, standards would reflect practices that are ultimately linked with better learning, but the evaluation process takes this linkage as given. Second, the evaluations are based on observations of one pre-announced class by a school administrator. Teacher (and student) behavior during this visit most likely differs from the norm, and administrator decisions may be distorted by interactions outside the classroom. Finally, the evaluations do not provide adequate feedback for teachers to improve their performance. Presumably, many (if not all) teachers could benefit from constructive recommendations on their instructional practices, but the current evaluation simply identifies 2 percent of teachers who are unsatisfactory and provides little information to the other 98 percent of teachers.

Evidence for Using Value-Added Assessments in Teacher Evaluation

Over the past decade, numerous studies have used longitudinal student level data to estimate the contribution of teachers to student learning (Rivkin, Hanushek, and Kain, 2005; Gordon, Kane,

and Staiger, 2006; Harris and Sass, 2006; Aaronson, Barrow, and Sander, 2007; Clotfelter, Ladd, Vigdor, 2007; Koedel and Betts, 2007; Jacob and Lefgren, 2008; Kane and Staiger, 2008; Kane, Rockoff, and Staiger, 2008; Buddin and Zamarro, 2009; McCaffrey, Sass, Lockwood, and Mihaly, 2009; Kane, McCaffrey, Miller, and Staiger, 2013). These studies have relied on value-added methods that isolate teacher contributions to student outcomes by estimating the effects of teachers on student achievement conditional on prior year test score and student-level measures of student demographics and background. The value-added approach relies on teacher "output" as measured by improvements in student test scores. This approach is a sharp departure from orthodox measures of teacher quality that have relied on teacher preparation and training (e.g., education level, experience, or subject matter knowledge) and occasional classroom observations by a school administrator.

Value-added researchers typically find wide variability in teacher effects suggesting that some teachers may be much more effective than others at improving student achievement. Some findings are common across most studies.

- **Experience**. New teachers are typically less effective than others, but teacher effects vary little with experience after the first year or two of teaching experience (Harris and Sass, 2006; Kane, Rockoff, and Staiger, 2008; Buddin and Zamarro, 2009).
- **Advanced Degrees**. Teachers with master's degrees have similar effects to teachers with only bachelor's degrees (Aaronson, Barrow, and Sander, 2007; Koedel and Betts, 2007; Buddin and Zamarro, 2009).
- **Certification**. Teachers with alternative certification are often just as effective at improving test scores as teachers certified through traditional programs (Kane, Rockoff, and Staiger, 2008; Sass, 2011).
- **Distribution of Teacher Value-Added Scores**. High value-added teachers are widely distributed across schools and not concentrated in a few schools. After controlling for prior achievement, teachers in intercity schools often perform as well or better than their counterparts in more wealthy suburban schools. Teacher value-added scores vary more within schools than across schools (Rivkin, Hanushek, and Kain, 2005; Buddin and Zamarro, 2009).

Most criticism of value-added approaches has focused on concerns about ranking individual teacher performance. Several studies have raised concerns about various aspects of value-added assessment. These criticisms suggest that value-added measures may provide little information regarding the effectiveness of a teacher, or misleading information as to the affect a teacher may have on his or her students. However, there is a substantial amount of research that provides credible validity evidence to support the appropriate use of value-added measures. The remainder of this section discusses important criticisms of value-added measures and reports research evidence that addresses these criticisms.

Accuracy of Achievement Tests. Student achievement test are incomplete measures of student knowledge and even the best tests measure achievement with some error (Haertel, 2013; Baker et al., 2010). Despite the presence of measurement error in state tests, researchers have found similar value-added results when using separate tests. For instance, Kane, McCaffrey, Miller, and Staiger (2013) compared teachers' value-added estimates from state tests with those from separate math and English assessments that measured higher-order skills. Some critics have argued that teacher effects derived from state tests are misleading because these scores reflect "teaching to the test" and not a deeper level of student learning (Baker et al., 2010). The researchers found that student scores were highly correlated on the two sets of tests, so teachers that had high value-added scores on the state test in math or English were likely to have high value-added scores on the higher-order tests as well.

In addition to measurement error, there may be test ceiling effects where the student is not able to demonstrate growth due to the difficulty level of the test which would affect value-added measures. However, with the exception of some state minimum competency tests, researchers have not found ceiling effects that influence the value-added estimates (Koedel and Betts, 2009).

Poor Contextual Controls for Classroom Composition and Student Sorting. A second criticism of value-added methods is that most studies rely on district or state administrative data and have few controls for the mix of students assigned to an individual teacher (Haertel, 2013; Baker et al., 2010). For example, these control variables are often limited to gender, race/ethnicity, free/reduced lunch eligibility, English learner status, and special education status. With limited controls, researchers are unable to adjust estimates for the possible sorting of students into classrooms (Rothstein, 2008; Rothstein, 2009). If students are nonrandomly

assigned to classrooms, then teacher value-added scores may reflect nuances of the sorting mechanism instead of differences in actual teacher effectiveness. If some teachers are assigned "better" students than others, then they have an unfair advantage in the teacher ranking. Since value-added controls for prior achievement, "better" in this context is not simply high achieving students, but rather students with more potential for improvement in a given year.

In practice, the contextual controls and student sorting have not created large distortions in the value-added estimates. Chetty, Freedman, and Rockoff (2013) examined potential biases in estimated teacher effects due to the sorting of students into classrooms. For example, administrative data have weak measures of family socioeconomic status (SES). If high SES student were disproportionately concentrated with some teachers, then these teachers might appear more effective than others simply because of the selection of students into their classrooms. Chetty, Freedman, and Rockoff tested this sorting bias by comparing teacher effects from traditional administrative data versus data that included richer controls for family SES. The family characteristics included parental marital status, family income, mother's age at student's birth, and indicators for parental contributions to a 401(k) and home ownership. The researchers found that the absence of the family-level variables (e.g., marital status and income) in traditional estimates of teacher effects had little effect on those estimates. They argue that the bias in the teacher effects is small, because the effects of these family characteristics are implicitly included in the traditional models through the controls for lagged test scores.

Chetty, Freedman, and Rockoff (2013) also examined whether estimated teacher effects were consistent with changes in grade-level test scores as teachers moved from school to school. If the value-added methodology accurately captures persistent differences in teacher effectiveness, then the movement of a high-quality teacher from one school to another should have a predictable effect on achievement at both the old and new school. For example, when a high-quality fourth grade teacher leaves a school, the grade-level gains should fall at the school that the teacher leaves and should increase at the school that the teacher enters. Indeed, student test scores moved as predicted when teachers moved from school to school and provided further evidence that estimated teacher effects represent a persistent and real underlying difference in teacher effectiveness.

The landmark MET Project provides a substantial counter to the criticism of non-random student assignment to teachers (MET Project, 2012; Kane, McCaffrey, Miller, and Staiger, 2013). This massive study was conducted across six large metropolitan school districts. Teacher effectiveness was measured both through multiple classroom evaluations by trained observers and through value-added techniques. Teacher value-added scores were measured for an initial period and compared with estimates following the random assignment of students to teachers. Students were randomly assigned to teachers, so observed teacher effectiveness was not confounded by the types of students assigned to different teachers.

The researchers found that student sorting had little effect on value-added estimates of teacher effectiveness. Teacher rankings before and after random assignment were highly correlated with one another. This evidence suggests that student-level controls in value-added models may be adequate controls for differences in students assigned to individual teachers.

Stability of Value-Added Scores. Another criticism of value-added measures is that value-added estimates may vary from year to year as a teacher's classes change or as teacher effectiveness varies over time (Baker et al., 2010). If so, then value-added estimates would provide limited insight into which teachers were most effective or what teacher practices were most effective.

Several empirical studies show that value-added scores are relatively unstable from year to year, but the studies also find persistent, stable teacher effects when teachers are observed across multiple years and classes (Koedel and Betts, 2007; Buddin, McCaffrey, Kirby, and Xia, 2007; McCaffrey, Sass, Lockwood, and Mihaly, 2009). The instability of year-by-year estimates reflects measurement error in the tests as well as the small number of students taught by teachers in a given year. For example, elementary school classes often contain only 20 to 30 students, so teacher effects may be unduly affected by test results for a small number of students. The evidence suggests that this instability is sharply reduced if the teacher effects are based on even two or three years of data.

Relation to Other Teacher Quality Measures. Haertel (2013) criticizes value-added methods, because they provide no direct indication of why some teachers are more effective than others or

how individual teachers could improve. In contrast, classroom observation and teacher pedagogy approaches provide better hands-on recommendations for improving instruction.

While value-added studies have no information on classroom practices, a few recent studies have shown linkages between value-added estimates and teaching practices. This linkage suggests that low value-added teachers may be able to improve their effectiveness and value-added scores by implementing better teaching practices.

Kane, Taylor, Tyler, and Wooten (2010) combined data on teacher value-added scores in Cincinnati with multiple classroom evaluations of each teacher by trained evaluators.[1] The study found that value-added measures and classroom observations were highly correlated and that improvements in classroom practices were likely to improve student achievement growth.

Grossman et al. (2013) examined the relationship between instructional practices and value-added assessments for middle school English Language Arts teachers. The study relied on trained evaluators observing instructional practices using an observational protocol. They find that high value-added teachers employ much more effective instructional practices than low value-added teachers.

The MET Project included detailed classroom evaluations by multiple trained evaluators as well as value-added estimation. The results showed substantial variability of teachers by different evaluators even with detailed evaluation protocols. The classroom evaluations under various protocols were highly correlated with teacher value-added scores on both state and higher-order skills tests. This study reinforces the notion that value-added estimates reflect underlying differences in instructional practice, and the estimates are consistent with recently developed instructional protocols.

The MET Project concluded that teacher evaluation should include value-added estimates, multiple and detailed classroom observations, and student survey information. The value-added estimates provide a cost-efficient method for differentiating less from more effective teachers; while the more detailed classroom observations can provide teachers with meaningful feedback to improve their instructional practices.

[1] The evaluations were based on the framework developed by Charlotte Danielson in *Enhancing Professional Practice: A Framework for Teaching.*

Predictive Evidence for Value-Added Assessments. In addition to other findings, recent predictive evidence from value-added assessment provides further support for the validity of the approach. Chetty, Freedman, and Rockoff (2013) relied on twenty years of data on students and teachers in grades 3 through 8 for a large metropolitan school district. The researchers constructed value-added estimates of teacher effects and tracked the long-term effects of teachers on the adult outcomes of their students. They found that students with high value-added primary school teachers were more likely to attend college, earn high wages as adults, live in higher SES neighborhoods, and have higher savings rates. These persistent effects of high value-added teachers on their students provide evidence in support of the validity of the value-added measures.

Glazerman et al. (2013) provided additional evidence on the persistence of teacher effects as teachers move from school to school. The study investigated the use of financial incentives to encourage high value-added teachers (top 20 percent) to volunteer for an assignment in low-achieving schools. Vacant teaching positions were randomly assigned to be filled by a high value-added teacher with a $20,000 incentive (the treatment group) or by another teacher through normal hiring practices. The high value-added teachers had positive effects on elementary test scores in the low-achieving schools relative to the group of control teachers.[2]

Dee and Wyckoff (2013) analyzed the effects of a Washington, D.C. teacher evaluation system that was based on a combination of structural classroom evaluations and value-added assessments. They found that the evaluation system encouraged low-performing teachers to voluntarily leave district positions, and those that remained made large student achievement gains in their classrooms. In addition, financial incentives for high-performing teachers were associated with high classroom achievement gains. This evidence suggests that value-added measures may be an important component of teacher evaluation that could substantially improve student achievement.

2. Legal Challenges

[2] The study did not find significant differences in middle school student achievement between the treatment and control groups. The authors argue that the insignificance of the middle school results may reflect small sample sizes in the middle school analysis or district-specific issues in the study districts.

The value-added research illustrates the promise and the potential problems associated with the use of value-added assessment as part of teacher evaluation reform. The promise of value-added assessment is the ability to accurately differentiate the effectiveness of teachers so that low performers can receive additional resources for improvement. In addition, administrators and teachers might learn from the instructional practices of high performers or even use these teachers as mentors for less effective teachers. One of the problems of incorporating value-added methods into teacher evaluation is how it can be applied district or statewide for all instructional staff.

There have been two lawsuits filed since 2012 that illustrate the issues. In California, *Vergara v. State of California*, No. BC484642 (Cal. Super. Ct.) highlights the promise of value-added assessment for teacher evaluation through its challenge to portions of California's Education Code. The plaintiffs want districts to improve measures of teacher effectiveness, which would include the use of value-added scores, when making employment, retention and termination decisions about teachers. In Florida, *Cook v. Stewart*, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.) illustrates some of the challenges that can occur when using value-added measures for all teachers. The issue in *Cook* is a reform to the Florida teacher evaluation system, the "Student Success Act," which requires a percentage of employee performance evaluations to include data of student learning growth. The next section will detail the two lawsuits, highlighting the evaluation system at issue and the claims of the plaintiffs.

Vergara v. State of California

The lawsuit *Vergara v. State of California*, filed in May 2012, alleges that portions of the California Education Code related to teacher tenure, dismissal, and layoffs have resulted in the continued employment of grossly ineffective teachers and in the subsequent denying of equal access to the opportunity to receive a meaningful education, which is a right guaranteed through provisions of the California state constitution. Although the lawsuit affects only California schools, other state constitutions have similar provisions, so a favorable ruling for the plaintiffs would have large implications for litigation in other jurisdictions.

The California plaintiffs are nine school-aged children, ranging in age from seven years old to fifteen years old, are of various backgrounds, and are allegedly at substantial risk of being

assigned a grossly ineffective teacher.[3] All but one of the plaintiffs attends a traditional public school, with the other attending a public charter school. Likewise, all but one of the plaintiffs has been assigned to at least one grossly ineffective teacher. In court filings, the plaintiffs detail some of their classroom experiences. For instance, one plaintiff claimed that one of her teachers permitted students to smoke marijuana during class and made statements that Latino students "would 'never graduate' and would instead 'clean houses for a living'" (Plaintiff's Opposition to Motion for Summary Judgment, Vergara v. State of California, No. BC484642 (Cal. Super. Ct.)). Another plaintiff asserted that she was assigned multiple grossly ineffective teachers in early elementary school and was not able to read by third grade. Once transferred to a public charter school, she succeeded academically and attained proficiency.

The plaintiffs contend that three sections of the California Education Code are partially responsible for their assignment to grossly ineffective teachers. The first section is the permanent employee statute (Cal Ed. Code Section 44929.21(b)), which makes a probationary employee permanent after two years. The decision to make a probationary teacher permanent must be made by March 15 of the second year teaching. If the district does not make a decision, the teacher is automatically reelected and made permanent. Because the decision must be made in the spring of the teacher's second year, administrators are only able to rely on a year and a half of observational data and likely only a year[4] of student performance data.

The second challenged section is the dismissal statute (Cal. Ed. Code Sections 44934, 44938(b)(1) and (2)). Plaintiffs contend that the statutes make dismissal nearly "impossible." To dismiss a permanent teacher, the district must follow a number of due process steps that are required by the statute. The steps include: (1) written notice specifying instances of behavior with particularity to provide the teacher with an opportunity to correct the behavior; (2) at least 90 days to correct performance; (3) after 90 days, file a written statement of charges, provided that it is not in the final one-fourth of the school year, otherwise the filing must wait until the following school year; (4) 30 days for the teacher to request a hearing; (5) 60 days for the hearing to commence, unless extended for good cause; and (6) a written decision by the

[3] An issue at trial is how "grossly ineffective teachers" are identified and if plaintiffs were actually assigned to "grossly ineffective teachers." Plaintiffs assert that value-added measures can be used to identify effective teaching, but defendants refer to it as a "'flawed methodology'" (Fensterwald, 2014).
[4] California administers their statewide assessment in the late spring.

administrative panel containing findings of fact, determinations of issues, and a disposition which is final unless appealed. Appeals can then be made to the California Superior Court and then to the California Court of Appeals. If the district loses, it must pay the expenses for the dismissal hearing, expenses incurred by the administrative panel, and the teacher's attorney's fees. In the Los Angeles Unified School District, the dismissal process costs between $284,932 and $404, 806 per teacher and can take 4 to 5 years (Plaintiffs' Opposition to Summary Judgment, Vergara v. State of California, No. BC484642 (Cal. Super. Ct.)). California employs about 275,000 teachers annually, but only 91 permanent teachers have been dismissed for cause statewide since 2003, and of those only 19 of those were dismissed for unsatisfactory performance (Plaintiffs' Opposition to Petition for Writ of Mandate, Vergara v. State of California, No. BC484642 (Cal. Super. Ct.)).

The third challenged section is the last-in-first-out statute (Cal. Ed. Code Section 44955). As the name implies, last-in-first-out means that layoffs are conducted by seniority, without any consideration of teacher competency. With the exception of some forms of specialized training, districts do not have the discretion to define "competency" in a way that would permit them to lay off a less effective, more senior teacher so that they could retain a more effective, junior teacher.[5]

The plaintiffs contend that "the State of California is *knowingly* forcing school districts to place some of their students, year after year, in classrooms with teachers who hinder the students' academic progress and cause severe and lasting harm." (Plaintiffs' Opposition to Motion for Summary Judgment, Vergara v. State of California, No. BC484642 (Cal. Super. Ct.)). Through the assignment of students to grossly inefficient teachers, the plaintiffs allege a violation of equal protection, infringing on the fundamental right to a public education under California's state constitution.

[5] In *Reed v. California* (Feb. 8, 2011, No. BC432420), the California Superior Court granted a preliminary injunction to enjoin the Los Angeles Unified School District from conducting seniority-based layoffs at three district middle schools where the reduction in force created a large number vacancies the prior year and would likely create a large number of vacancies during the 2009-2010 school year. The plaintiffs and the district entered into a consent decree that would minimize layoffs at targeted schools, which would be affected the greatest by seniority-based layoffs. The consent decree was later challenged by the teachers union and declared unenforceable. *Reed v. United Teachers Los Angeles*, 208 Cal. App. 4th 322 (Aug. 10, 2012). For more information see Jared S. Buszin, Beyond School Finance: Refocusing Education Reform Litigation to Realize the Deferred Dream of Education Equality and Adequacy, Emory Law Journal, 62 Emory L.J. 1613 (2013) .

In the request for relief, plaintiffs request a permanent injunction enjoining defendants from implementing any system of teacher employment, retention, and dismissal that is substantially similar to the present framework that provide teachers greater protections against dismissal than the rights applicable to other California state employees or prevents school administrators from "meaningfully considering teacher effectiveness when making employment, retention and termination decisions about teachers." (Plaintiffs' Complaint, Vergara v. State of California, No. BC484642 (Cal. Super. Ct.)). Although the complaint does not specify what the plaintiffs mean by "teacher effectiveness," on plaintiff's attorneys website, Students Matter, they provide further information stating that "[w]hile Students Matter does not prescribe any particular evaluation method, we do believe that any evaluation system should take into account multiple measures and meaningfully include some measurement of student performance. There are many groups and experts out there doing research on evaluations and piloting fair and accurate evaluation programs." (Students Matter, n.d.)

In sum, *Vegara* is seeking to require the state of California to make changes to teacher employment, retention, and dismissal practices that have the effect of keeping grossly ineffective teachers in the classroom. A key component of reforming the California system is to implement a meaningful teacher evaluation system—that would likely include student performance measures—which could be used for employment decisions.

Cook v. Stewart (previously *Cook v. Bennett*)

While *Vegara* makes the argument that better teacher evaluation systems using student performance data should be implemented and used for employment decisions, the Florida lawsuit *Cook v. Stewart* illustrates some of the complications when student performance measures are used for all instructional staff as part of performance evaluations.

Cook was filed in April 2013 and is a challenge to the "Student Success Act," (Senate Bill 736, codified as chapter 1012, Florida Statutes.) The Student Success Act requires annual evaluations of all instructional employees. Instructional employees include classroom teachers and other employees who provide direct support in the learning process of students outside of a regular classroom such as guidance counselors and librarians. (Fla Stat. 1012.01(2)). The evaluations must be based, in part, on student learning growth as measured by the statewide assessments or,

for subjects and grade levels not measured by statewide assessments, by school district assessments (Fla. Stat. 1012.34(3)(a)(1)). Districts have until the 2014-2015 school year to adopt assessments for each offered course (Fla. Stat. 1008.22(6)(b)). For districts that have not adopted a district-created assessment or approved-student learning growth measure for non-tested grades and subjects, "measurable learning targets must be established based upon the goals of the school improvement plan and approved by the school principal. "(Fla. Stat. 1012.34(7)(e)). Further, "a district school superintendent may assign to instructional personnel in an instructional team the student learning growth of the instructional team's students on statewide assessments." (Fla. Stat. 1012.34(7)(e)). The Act does not define what constitutes an instructional team.

As originally enacted, the statute did not specify that the growth must be of the teacher's own students. The statute was amended to clarify that the evaluations "must be based upon the performance of students assigned to [the instructional personnel's] classrooms or schools." (Fla. Stat. 1012.34(3)).

The amount of the evaluation based on student learning growth varies depending on the type of employee and the number of years of available data. (Fla Stat. 1012.34(3)(a)). Table 1 illustrates how the percentage of the evaluation is reduced by 10 percentage points when there are less than three years of data available. Likewise, less of the performance evaluation is based on student learning growth for other instructional employees who are not classroom teachers. The statute does not, however, reduce the amount of the evaluation based on student growth for classroom teachers who teach in non-tested grades and subjects.

Table 1. Percentage of Evaluation Based on Student Learning Growth Based on Employee Type and Years of Data

	Teacher Evaluation Percentage	
	3 or More Years of Data	Less than 3 Years of Data
Classroom Teacher	No less than 50%	No less than 40%
Other Instructional Employee	No less than 30%	No less than 20%

The lack of a distinction among classroom teachers who teach grades and subjects tested through the statewide assessment and those who are not is the primary basis for the lawsuit.[6] The plaintiffs in *Cook* are primarily teachers in non-tested subjects such as art, music, and health, and they are challenging the application of statewide reading and/or math scores to their subject alleging that it is a violation of plaintiffs' substantive due process and equal protection rights under the Fourteenth Amendment of the United States Constitution. [7]

Regarding due process they claim a state cannot impose punishments or burdens for individuals for "actions over which they have no responsibility or ability to control." (First Amended Complaint, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)). According to the plaintiffs, their students' reading and mathematics scores are beyond their control as their courses are not designed to teach reading and/or mathematics.[8] Their expert witness, Edward Haertel, described it as an "'extreme mismatch between the curriculum and the test used to assess student learning'" such that the value-added scores "'provide virtually no information about the effectiveness of the teaching.'" (Plaintiffs' Motion for Summary Judgment, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)). Further, they contend that there is no evidence of reliability or validity for using student test scores for teachers in non-tested subject areas (Plaintiffs' Motion for Summary Judgment, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)).

For the equal protection challenge, they argue that the Student Success Act creates separate classes of teachers in Florida: "those whose evaluations are based on student growth data for students assigned to the teacher in the subjects taught by the teacher, and those whose evaluations are based on student growth data for students and/or subjects they do not teach."

---

[6] Another basis for the complaint was that under the Florida system, some teachers were evaluated based on test scores of students that they did not teach. For instance, plaintiff Cook was a first grade teacher who started teaching at Irby Elementary in 2011. Irby Elementary only serves students in preschool through second grade, none of which take the statewide assessment. To include student learning growth for teachers in Irby Elementary, the school district used the test scores for students in fourth and fifth grade at Alachua Elementary School, which is the school the Irby Elementary students attend after second grade. The amendment to the Student Success Act would likely ensure that this type of evaluation will not occur in the future.

[7] Plaintiffs also contend that they have suffered emotional distress, reputational harm, and have suffered other injuries, including the potential loss of employment, due to the evaluation system.

[8] The State Defendants argue that the legislature could have rationally believed that student learning in one subject could carry over to another subject (State Defendants' Motion to Dismiss, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)). Further, it should be noted that the problem is temporary, as districts must adopt assessments for all offered courses by the 2014-2015 school year. Fla. Stat. 1008.22(6)(b); Fla. Stat. 1012.34(7)(e)).

(First Amended Complaint, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)). Teachers are then classified as "highly effective," "effective," "needs improvement," or "unsatisfactory" based on the use of the test scores that plaintiffs contend were not designed to measure teacher performance in non-tested subjects (Plaintiffs' Motion for Summary Judgment, Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)). Plaintiffs assert that the creation of the classes is a violation of equal protection arguing there is no rational justification for basing the evaluations for teachers of non-tested grades and subjects on student performance for classes they do not teach. In the request for relief, plaintiffs are seeking for the entire Student Success Act to be declared unconstitutional under the Fourteenth Amendment of the Constitution for all employees, not just those in untested grades and subjects, and permanently enjoining the implementation or enforcement of the Act.

Discussion

As argued in *Vergara*, access to public education is a fundamental right under California's state constitution[9] and student assignment to grossly ineffective teachers interferes with that right to a public education. Value added methods are a way to defensibly demonstrate which teachers are essentially denying that fundamental right. The implication for teacher evaluation systems is that the system should be able to accurately differentiate the effective teachers from the ineffective teachers. Many evaluation systems often do not do this. Instead, by identifying every teacher as "satisfactory," we are unable to identify which teachers are truly ineffective and in need either of remediation or to be removed from the classroom. Thus, it is necessary to ensure that there is accurate differentiation among teachers.

Value-added measures can provide such differentiation, but steps must be taken to ensure their accuracy. By following some of the general guidelines from the research literature related to the appropriate use of value added measures, a state or district could better defend the use of the measures as part of a teacher evaluation system. Specifically, for teachers in self-contained classrooms (e.g., most elementary school teachers) who only have 20 or 25 students per year there must generally be multiple years worth of data. Three years is ideal, as it improves year-to-year stability in estimates, but two years may also be sufficient (Koedel and Betts, 2007; Buddin, McCaffrey, Kirby, and Xia, 2007; McCaffrey, Sass, Lockwood, and Mihaly, 2009). Second,

[9] Most other states have similar provisions (Wilkins, 2005).

value-added assessment should not be the sole measure of the evaluation (AERA/APA/NCME, 1999). Instead, other measures such as classroom observations and parent and student surveys should be included, and the weighting of each measure should be balanced so that one measure does not dominate the evaluation (MET, 2013).

Although the general value-added guidelines apply for all teachers, the larger issue that the *Cook* case raises is what should be done for teachers in non-tested grades and subjects. The districts in *Cook* used subject area student achievement scores in a statewide test to evaluate teachers in non-tested grades and subjects. The problem in doing so is that there is a tenuous and untested relationship between the teacher's subject matter and the assigned tested subject. For example, the curriculum in art and music may inherently provide little opportunity for teachers in these subjects to improve the reading and math skills of their students. On the other hand, the Common Core State Standards literacy standards that are applicable in "history/social studies, science, and technical subjects" (Common Core State Standards, n.d.) The literacy standards are meant to supplement, not replace, the content standards in those subject areas. States and districts that have adopted the literacy standards for these subject areas have a stronger rationale for applying statewide reading scores to other subjects than those that do not explicitly require non-reading teachers to teach literacy. The courts may look to the state content standards and any written district curriculum to determine if there is a rational basis for the assignment of test scores from other subject areas.

As mentioned previously, the application of student test scores from other subject areas is only temporary in Florida. However, requiring districts to adopt or create assessments for currently non-tested grades and subjects will likely present new challenges for districts. Creating assessments is challenging. In addition to developing the content, districts would also need to collect evidence of the reliability and validity of the test scores and be able to present this evidence if there was a legal challenge. It is unlikely that a Florida district would have the resources to do all of the work required for a psychometrically defensible assessment. To address the capacity issue, the state is providing grants to aid in the creation of test item development for hard-to-measure content areas, such as physical education and the performing arts, and is providing assessment development training to districts (Florida Department of Education, 2013).

However, because these are hard-to-measure content areas they may be more heavily scrutinized and need additional validity evidence.

Summary

Teachers are an essential part of student learning. They have long-term, substantial, and substantive impacts on students. As the plaintiffs in *Vergara* testified, there is a need to ensure that students receive an appropriate education and are not assigned to grossly ineffective teachers. To ensure this, we must first be able to accurately identify a grossly ineffective teacher with some measure of objectivity, and, despite their criticisms, value-added measures—when used in conjunction with other methods to give teachers meaningful feedback to improve instruction—provide an effective and efficient way of doing so.

As states begin to change teacher evaluation systems to incorporate the use of student performance data, there likely will be additional legal challenges related to how to value-added measures are incorporated, particularly for teachers in untested grades and subjects. As the *Cook* case in Florida illustrates, states and districts must have rational reasons for attributing student test scores to teachers. The rational basis standards that courts would likely employ only requires that there be a rational relationship to the legitimate governmental interest involved, and the court gives deference to the government. When the teacher is responsible for a subject that is directly related or where the content has been incorporated into the teacher's required curriculum there is an obvious relationship between the uses of value-added measures as part of teacher evaluation systems. States and districts must have a more thoughtful rationale when there is less of an obvious connection, but as long as there is a rationale the court generally will uphold the law. Further, as districts begin to develop assessments for currently untested grades and subjects, they will be responsible for also collecting validity evidence to support their use and potentially modify their systems in light of the validity evidence.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labor Economics, 24*(1), 95-135.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., & Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*(4), 416-440.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics, 66*, 103-115.

Buddin, R., McCaffrey, D., Kirby, S., & Xia, N. (2007). *Merit pay for Florida: design and implementation issues*. WR-508-FEA, Santa Monica, CA: RAND.

Chetty, R., Friedman, J., & Rockoff, R. (2013). *Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates.* NBER Working Paper No. 19423, Cambridge, MA: National Bureau of Economic Research.

Chetty, R., Friedman, J., & Rockoff, R. (2013). *Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood.* NBER Working Paper No. 19424, Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). *How and why do teacher credentials matter for student achievement?* NBER Working Paper No. 12828, Cambridge, MA: National Bureau of Economic Research.

Common Core State Standards (n.d.). *English language arts standards*. Retrieved from http://www.corestandards.org/ela-literacy.

Cook v. Stewart, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.).

Dee, T., & Wyckoff, J. (2013). *Incentive, selection, and teacher performance: evidence from IMPACT.* NBER Working Paper No. 19529, Cambridge, MA: National Bureau of Economic Research.

Duncan, A. (2013, June 18). Letter to the Chief State School Officers. Retrieved from U.S. Department of Education website: http://www2.ed.gov/policy/elsec/guid/secletter/130618.html.

Fensterwald, J. (2014, Jan. 28). In landmark trial, both sides debate whether teacher protection law fails students, *The Hechinger Report*. Retrieved from http://hechingerreport.org/content/in-landmark-trial-both-sides-debate-whether-teacher-protection-laws-fail-students_14576/.

Florida Department of Education (2013), *Developing local assessments*. Retrieved from http://www.fldoe.org/asp/k12memo/pdf/DevelopingLocalAssessments.pdf.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G.J. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

Glazerman, S., Protik, A., Teh, B., Bruch, J., & Max, J. (2013). *Transfer incentives for high-performing teachers: final results from a multisite randomized experiment.* Washington, DC: U.S. Department of Education.

Goldhaber, D., & Loeb, S. (2013). What Do We Know About the Tradeoffs Associated with Teacher Misclassification in High Stakes Personnel Decisions? *The Carnegie Knowledge Network.*

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job.* Brookings Institution Discussion Paper 2006-1, Washington, DC: The Brookings Institution.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: the relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445-470.

Haertel, E. (2013). *Reliability and validity of inferences about teachers based on student test scores.* Princeton, NJ: Educational Testing Service.

Harris, D., & Sass, T. (2006). *The effects of teacher training on teacher value-added.* Working paper, Florida State University.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: an experimental evaluation.* NBER Working Paper No. 14607, Cambridge, MA: National Bureau of Economic Research.

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T., Rockoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27,* 615-631.

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2010). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46*(3), 587-613.

Koedel, C., & Betts, J. (2007). *Re-examining the role of teacher quality in the educational production function.* Working paper, University of California, San Diego.

Koedel, C., & Betts, J. (2009). *Value-added to what? How a ceiling in the testing instrument influences value-added estimation*. NBER Working Paper No. 14778, Cambridge, MA: National Bureau of Economic Research.

McCaffrey, D., Sass, T., Lockwood, J., & Mihaly, K. (2009). The inter-temporal variability of teacher effects estimates. *Education Finance and Policy, 4*(4), 572–606.

MET Project, 2012. *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. MET Project Research Paper, Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

MET Project, 2013. *Feedback for better teaching: Nine principles for using measures of effective teaching.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf.

National Council on Teacher Quality (2004). *Increasing the odds: How good policies can yield better teachers.* Washington, DC: National Center on Teacher Quality.

Newton, S. (2004). *Stull evaluations and student performance.* Los Angeles Unified School District, Planning, Assessment and Research Division Publication No. 186.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement, *Econometrica, 73*(2), 417-458.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214.

Rothstein, R., Ladd, H.F., Ravitch, D., Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Linn, R.L., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Sass, T. (2011). *Certification requirements and teacher quality: a comparison of alternative routes to teaching.* Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, American Institutes for Research.

Students Matter (n.d.). *Evaluations*. Retrieved from http://studentsmatter.org/our-case/vergara-v-california-case-summary/evaluations/.

U.S. Department of Education. (2013, June 14). *ESEA flexibility state-by-state timeline implementation chart*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/eseaflexstchart614.doc.

Vergara v. State of California, No. BC484642 (Cal. Super. Ct.).

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn: The New Teacher Project. Retrieved from http://widgeteffect.org.

Wilkins, B. (2005). Should public education be a federal fundamental right? *Brigham Young University Education & Law Journal*, 2005, 261.