

# The Legality of Using School-Wide Growth Measures in Teacher Evaluation Systems

MICHELLE CROFT, PHD, JD, AND RICHARD BUDDIN, PHD

Since 2009, more than two-thirds of states have enacted legislation to reform the systems by which public school teachers are evaluated.<sup>1</sup> A key change in the new systems is the inclusion of student growth measures as an evaluation criterion. These measures use student test scores to determine how much academic progress a student has made in a year.<sup>2</sup>

The use of objective measures of student achievement in teacher evaluation is a large and rapid shift in direction for states. In 2010, 16 states included such measures in teacher evaluations; by September 2013 the number of these states had more than doubled, to 35.<sup>3</sup> And the number is expected to increase to over 40 by the 2015–2016 school year, as states that were granted an Elementary and Secondary Education Act (ESEA) flexibility waiver adopt student growth measures for educator evaluations as a condition of receiving the waiver.<sup>4</sup>

The interest in student growth measures for teacher evaluation is driven by three factors. First, research evidence has emerged that some teachers are more effective than others at improving student performance in a given subject and grade.<sup>5</sup> Although this research has been controversial, many researchers and educators believe it suggests that student growth measures

offer an innovative and objective method for assessing teaching quality. Second, in 2011, the federal government began the Race to the Top grant program to fund education reform initiatives. One goal of Race to the Top was to encourage states and districts to measure student growth and use the measures to improve instruction.<sup>6</sup> Finally, because nearly all states would likely fall short of meeting the No Child Left Behind (NCLB) goal of 100 percent of students meeting proficiency targets in reading and math by 2014,<sup>7</sup> the US Department of Education initiated ESEA flexibility waivers for states in 2011. The waivers required that states develop comprehensive plans to improve student outcomes through reforms, including better teacher evaluation systems.<sup>8</sup>

Historically, student growth had rarely been taken into account in evaluating teachers even in core subjects, such as English language arts (ELA) and mathematics, where students are frequently tested. Early research using value-added methods, of which student growth measures are one, was often limited to these subjects<sup>9</sup> because value-added methods require test results in the same content area in continuous grades, and statewide testing typically occurs only in ELA and math in grades 3–8 and once in high school.<sup>10</sup> Despite the fact that research had been limited to ELA

**Michelle Croft** is a principal research associate in the Office of Policy, Advocacy, and Government Relations at ACT.

**Richard Buddin** is a principal research economist at ACT. His research focuses on statistical evaluations of education and workforce programs.

and math, both Race to the Top and the ESEA flexibility waivers required student achievement measures for all teachers,<sup>11</sup> and the new systems are following suit, covering not only ELA and math teachers but also teachers of subjects such as music and physical education, where students are not tested as regularly or at all.

As a result, because student growth measures are not available for all teachers, states and districts have had to seek other solutions in order to incorporate growth into the evaluations of all teachers. One popular but controversial approach is to use school-wide growth measures, which are typically based on the ELA or math scores for all the students in a school who took the statewide tests. Under this approach, the student growth component of a physical education teacher's evaluation, for example, would come from his or her students' scores on a statewide test that has nothing to do with physical education. Depending how school-wide growth measures are defined, the student growth component may also include students who are not even in the teacher's classes.<sup>12</sup>

The use of school-wide growth measures for non-tested grades or subjects, as this practice is called, has so far prompted two lawsuits, in Florida and Tennessee. This brief summarizes the legal cases and, inspired by the cases, provides recommendations on the appropriate use of school-wide growth measures.

### **Florida: *Cook v. Stewart***

*Cook* was filed in Florida's federal district court in April 2013 by teachers in non-tested subjects (e.g., art, music, and health) and challenged Florida's Student Success Act.<sup>13</sup> The Student Success Act requires annual evaluations of all instructional employees, which include classroom teachers and other employees who provide direct support to students (e.g., guidance counselors and

librarians).<sup>14</sup> The Student Success Act was phased in, so that initially districts could assign instructional personnel to teams and use school-wide growth measures in reading and mathematics with educators for whom subject-specific student growth measures were not available.<sup>15</sup>

The plaintiffs in *Cook* challenged the use of school-wide growth measures as violations of plaintiffs' substantive due process and equal protection rights under the Fourteenth Amendment of the United States Constitution.<sup>16</sup> Plaintiffs claimed that their students' reading and mathematics scores were beyond their control (and responsibility) given that their courses, which included art, music, and health, were not designed to teach reading or mathematics.<sup>17</sup> They also asserted that there was no evidence of reliability or validity to support the use of student test scores for teachers in non-tested subject areas.<sup>18</sup> Likewise, they argued that the Student Success Act created separate classes of teachers in Florida depending on where subject-specific student growth data was available. Plaintiffs asserted there is no rational justification for basing the evaluations for teachers in non-tested grades and subjects on performance by students and in subjects they do not teach.

### **Tennessee: *Wagner v. Haslam***

Tennessee's First to the Top Act of 2010 required that 50% of a teacher's evaluation consist of student achievement measures: 35% on student growth from the statewide Tennessee Value-Added Assessment System (TVAAS) tests in reading and math, and 15% on "other measures" of student achievement from a state-approved list.<sup>19</sup> Teachers without subject-specific TVAAS scores could use scores on other state-approved assessments, if available, or school-wide growth measures.

Similar to the *Cook* plaintiffs, the two plaintiffs in *Wagner* taught subjects that were not tested statewide.<sup>20</sup> Plaintiff Wagner

taught physical education for 26 years, and plaintiff Braeuner taught visual arts for six years.<sup>21</sup> Both plaintiffs received satisfactory evaluation scores meriting bonuses in 2011–2012 and 2012–2013—a level five (out of five) overall for Wagner and a level four (out of five) overall for Braeuner.<sup>22</sup> The school-wide TVAAS-based score for these years was a level five for both plaintiffs. In the 2013–2014 year, their school-wide TVAAS-based scores declined to a level one, and the overall evaluation scores for both plaintiffs decreased to a level three.<sup>23</sup> As a result, neither plaintiff received a bonus for 2013–2014<sup>24</sup> and, as a teacher at the end of her probationary status, Braeuner was not considered for tenure and would need to wait at least two years until she had an overall score of at least a level four for both of those years to qualify for tenure.<sup>25</sup> Both plaintiffs claimed that they suffered emotional distress, anxiety, diminished professional morale, and harm to their professional reputations.<sup>26</sup>

The *Wagner* plaintiffs contended that the evaluation system violated due process because they were evaluated on the basis of student test scores unrelated to the subjects they teach.<sup>27</sup> Plaintiffs claimed that the scores did not reflect their teaching ability because they are not licensed, employed, or required to provide instruction in reading or math.<sup>28</sup>

The *Wagner* plaintiffs also contended that the evaluation system was a violation of equal protection. They claimed that there was not a rational reason to classify teachers in non-tested subjects based on student test scores unrelated to courses they teach.<sup>29</sup> Plaintiffs contended that this "serves to undermine, rather than advance, the quality of Tennessee public schools."<sup>30</sup>

In the complaint, the *Wagner* plaintiffs also argued that the evaluation system did not effectively differentiate among educators using the school-wide growth measure, as

every teacher in non-tested grades and subjects received the same score on this measure.<sup>31</sup>

## Court Decisions

In both *Cook* and *Wagner*, the courts found that the use of school-wide growth measures meets the minimum rational basis standard in that the state has a legitimate interest in increasing student growth. Although both courts ruled against the plaintiffs, it is important to note that when evaluating constitutional challenges, the level of review (i.e., scrutiny) depends on the type of challenges being made. If the challenge involves a suspect classification (such as race or national origin) or a fundamental right (such as voting), *strict scrutiny* applies. With strict scrutiny, the burden is on the government to show that the challenged classification or right serves a compelling state interest and is necessary to serve that interest. If the challenge does not involve a suspect classification or fundamental right, as is the case with teacher evaluation systems, minimum *rational basis scrutiny* applies. This level of scrutiny is much lower than strict scrutiny, and with it the government only needs to show that the challenged classification is rationally related to a legitimate state interest.

In *Cook*, the court found it rational to believe that teachers may affect all students in their schools through influencing the school environment and inspiring other teachers.<sup>32</sup> In *Wagner*, the court highlighted multiple ways in which the state could have rationally believed that the use of school-wide growth measures in teacher evaluation may increase student performance: the quality of a teacher affects the school as a whole, and teachers may be more likely to participate in school-wide initiatives or to incorporate concepts from tested subjects into their own curriculum;<sup>33</sup> a good teacher may help students become better learners

in other classes;<sup>34</sup> or teachers may serve as a mentor to students other than those they directly teach.<sup>35</sup> The *Wagner* court highlighted that organization-wide criteria are used in other employment settings such as with salespeople, attorneys, or managers.<sup>36</sup> The *Wagner* court did not address whether the contested policies were efficient or wise but merely that they were a plausible way to improve student outcomes.

On the other hand, the judge in *Cook* said that the evaluation system was unfair and that he “would be hard-pressed to find anyone who would find this evaluation system fair to non-FCAT [Florida Comprehensive Achievement Test, the statewide assessment in ELA and math] teachers, let alone be willing to submit to a similar evaluation system.”<sup>37</sup> Similarly, the court in *Wagner* addressed the fairness issue by highlighting that unfair does not mean irrational.<sup>38</sup> The *Wagner* decision stated that “policymakers can make policies that address a policy issue in part and leave further refinement for later.”<sup>39</sup> The *Wagner* court acknowledged that teachers in tested subjects may be better off in the evaluation system than teachers in non-tested subjects, but the state and the State Board does not have to choose between using the individual value-added measures for all teachers or for none of them.<sup>40</sup> The State Board is able to act incrementally and use rational alternatives when the “best” option is not available.<sup>41</sup> The court also noted that the evaluations were based on other measures in addition to the school-wide growth measures.<sup>42</sup>

## Recommendations

Teachers are an essential part of student learning, and evaluations can serve as a useful tool to identify excellent teachers and provide necessary development programs and supports to other teachers when needed. Although two federal courts have upheld recent challenges to the use of

school-wide growth measures to evaluate teachers in non-tested subjects, both cases highlight concerns with ensuring that teacher evaluation systems adequately reflect teacher contributions to the academic growth of the students in their own subjects and classrooms and not (or not just) those of other teachers. Ideally, such systems should provide all teachers with a meaningful assessment of their strengths and weaknesses and encourage all teachers to improve their instructional quality.

To bring the new teacher evaluation systems closer to meeting this goal, we provide the following four recommendations:

1. **Limit the use of school-wide growth measures to teachers who have a reasonably direct effect on the measures.** For example, if a school instructional team pairs a social studies teacher and an ELA teacher to coordinate lessons and reinforce material from one another’s classes, evaluating the social studies teacher on their students’ statewide ELA value-added scores would be appropriate. Similarly, if the state’s standards or the district’s written curriculum requires literacy across subject areas, it may be appropriate to evaluate all teachers in a school based in part on growth measures derived from the scores of all students in the school on the statewide ELA test.
2. **Consider alternative measures of student growth for teachers in non-tested subjects.** These measures may include new assessments in these subjects, student portfolios of work in the classes,<sup>43</sup> or the other evidence that students met learning objectives for courses in non-tested subjects.<sup>44</sup> Although these alternatives potentially present additional concerns that would need to be addressed (such as comparability across classes, schools, or districts),<sup>45</sup> they would likely be more

appropriate in some cases than a school-wide measure because they are more directly relevant to the work teachers are doing with students in the non-tested grades and subjects.<sup>46</sup>

**3. Include teachers in the design and review of the evaluation system so that it is not only legal but equitable.**

Evaluation systems should be created and implemented so that they are not only legal but also equitable to teachers in both tested and non-tested grades and subjects. Teachers should be included in the initial design of the system, and in the process of reviewing and refining

a system after first or subsequent uses. If an evaluation system is not legal, it should certainly be scrapped and redeveloped, and if it is legal but not equitable, states and districts should take all reasonable steps, including incorporation of educator feedback, to make it an appropriate measure of quality for all teachers.

**4. Monitor the implementation of the evaluation system and include funding to support research on the effects of teacher evaluation systems.**

As states are experimenting different types of evaluation systems, research

is needed to understand not only the reliability and validity of the components of the evaluation system, but also the effects of the evaluation system. Ideally, the new systems would focus on identifying the effectiveness of various evaluation components at improving teaching quality as well as assessing how the components affect teacher retention and turnover. Ultimately, evaluation systems should be judged on the actual effectiveness of each component in practice.

**Notes**

- 1 National Conference of State Legislatures, *Evaluating Effective Teachers* (National Conference of State Legislatures, 2013), [www.ncsl.org/research/education/evaluating-effective-teachers635188303.aspx](http://www.ncsl.org/research/education/evaluating-effective-teachers635188303.aspx); Sara Mead, *Recent State Action on Teacher Effectiveness: What's in State Laws and Regulations?* (Bellwether Education, Washington, DC: 2012), [bellwethereducation.org/sites/default/files/legacy/2012/08/RSA-Teacher-Effectiveness.pdf](http://bellwethereducation.org/sites/default/files/legacy/2012/08/RSA-Teacher-Effectiveness.pdf); Race to the Top and the Elementary and Secondary Education Act (ESEA) Flexibility Waiver both required states to adopt teacher evaluation systems that included a measure of student growth. See generally, US Department of Education, *ESEA Flexibility Policy Document* (US Department of Education, June 7, 2012), [www2.ed.gov/policy/eseaflex/bie.pdf](http://www2.ed.gov/policy/eseaflex/bie.pdf); Race to the Top Fund: Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010, 74 Fed. Reg. 59844 (Nov. 18, 2009).
- 2 There are multiple types of student growth measures. Value-added methods that track individual test score gains to isolate a teacher effect are one of the more popular measures, but there are also measures such as assessments of progress in mastery of student learning objectives. See Michelle Croft & Richard Buddin, "Applying Value-Added Methods to Teachers in Untested Grades and Subjects," *Journal of Law and Education* 44 (2015):1–22.
- 3 Kathryn M. Doherty & Sandi Jacobs, *State of the States 2013: Connect the Dots: Using Evaluation of Teacher Effectiveness to Inform Policy and Practice* (Washington, DC: National Council on Teacher Quality, 2013), [www.nctq.org/dmsView/State\\_of\\_the\\_States\\_2013\\_Using\\_Teacher\\_Evaluations\\_NCTQ\\_Report](http://www.nctq.org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report). The report omitted the District of Columbia as a state entity and instead focused on the District of Columbia Public School District (DCPS).
- 4 As of September 2015, 43 states received ESEA Flexibility Waivers requiring teacher evaluation systems that include measures of student growth. US Department of Education, *ESEA Flexibility*, available at [www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html](http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html). Missouri and Kansas both received conditional waivers. Missouri has until 2015–2016 to implement the teacher evaluation system. Letter from Heather Rieman, Acting Assistant Secretary, US Department of Education, to Margie Vandeven, Commissioner of Education, Missouri State Department of Elementary and Secondary Education (June 23, 2015), available at [www2.ed.gov/policy/eseaflex/secretary-letters/morenewalltr.pdf](http://www2.ed.gov/policy/eseaflex/secretary-letters/morenewalltr.pdf). Kansas is required to use student growth for evaluations no later than the 2016–2017 school year. Letter from Heather Rieman, Acting Assistant Secretary, US Department of Education, to Brad Neuenswander, Interim Commissioner of Education, Kansas State Department of Education (June 23, 2015), available at [www2.ed.gov/policy/eseaflex/secretary-letters/ksrenewalltr.pdf](http://www2.ed.gov/policy/eseaflex/secretary-letters/ksrenewalltr.pdf).
- 5 For a summary of the research see Richard Buddin & Michelle Croft, *Recent Validity Evidence for Value-Added Measures of Teacher Performance* (Iowa City, IA: ACT, Inc., 2014), [www.act.org/research/policymakers/pdf/Measures-of-Teacher-Performance.pdf](http://www.act.org/research/policymakers/pdf/Measures-of-Teacher-Performance.pdf).
- 6 Department of Education Overview Information; Race to the Top Fund: Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010, 74 Fed. Reg. 59844 (Nov. 18, 2009).
- 7 Michele McNeil & Alyson Klein, "Obama Offers Waivers from Key Provisions of NCLB," *Education Week* (September 27, 2011), [www.edweek.org/ew/articles/2011/09/28/05waiver\\_ep.h31.html](http://www.edweek.org/ew/articles/2011/09/28/05waiver_ep.h31.html).
- 8 See generally *ESEA Flexibility Policy Document*, US Department of Education, (June 7, 2012), [www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html](http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html).

- 9 See Buddin & Croft, *Recent Validity Evidence*.
- 10 US Department of Education, *ESEA Flexibility*.
- 11 US Department of Education, *ESEA Flexibility*; Department of Education Overview Information; Race to the Top Fund: Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010, 74 Fed. Reg. 59844 (Nov. 18, 2009).
- 12 School-wide growth measures have also been used to evaluate overall ELA or math instruction within a school, (based on the average growth in math achievement at a school across all tested grades and classrooms). This measure emphasizes that achievement growth is a school-wide or team effort. In many cases, however, math or ELA growth is primarily related to instruction by a specific teacher matched with specific students for the school year (with some indirect spillover from math or ELA applications in other subjects). One problem with this type of school-wide average is that it masks difference in high and low performers. High performers are discouraged because they only receive partial credit, because performance gains are averaged with those of low performers. Teachers are nominally encouraged to contribute to the team effort, but this reward system implicitly encourages teachers, when placed in large groups, to “free ride” on the efforts of others and reduce their level of effort. A second limitation of using the school-wide average is that statewide testing often does not occur before grade 3. Using school-wide growth measures gives more weight to the impact of educators on students in grades 4 and higher.
- 13 Senate Bill 736, codified as chapter 1012, Florida Statutes.
- 14 Fla Stat. § 1012.01(2).
- 15 The Student Success Act did not define “instructional team.” The weight or percentage of student growth applied within the evaluation system varies depending on the number of years of data available and the type of employee. For teachers where there were no statewide tests available, 50% of their evaluation was based on student learning growth if there were three or more years of data available and 40% if less than three years of data. Fla Stat. § 1012.34(3)(a). Starting in the 2014–2015 school year, districts were required to have individual student growth measures for all instructional personnel. Fla. Stat. § 1012.34(7)(e).
- 16 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.
- 17 The State Defendants argued that the legislature could have rationally believed that student learning in one subject could carry over to another subject (State Defendants’ Motion to Dismiss). Further, it should be noted that the problem was temporary, as districts were required to adopt assessments for all offered courses by the 2014–2015 school year. Fla. Stat. 1008.22(6)(b); Fla. Stat. 1012.34(7)(e).
- 18 (Plaintiffs’ Motion for Summary Judgment, *Cook v. Stewart*, No. 1:13-cv-00072-MW-GRJ (D. N.D. Fl.)).
- 19 Tenn. Code Ann. § 302(d)(2)(A). There have been subsequent amendments to reduce the student growth percentages for teachers without individual student data. In 2015, only 30% of the evaluation criteria for those teachers may be based on student achievement, with 15% based on student growth. H.B. 108, 109th Gen. Assembly (Tenn. 2015) (amending Tenn. Code Ann. § 49-1-302(d)(2)(B)(vi)).
- 20 No. 3:2015-cv-00115, 2015 US Dist. Lexis 76443 (M.D. Tenn. 2015).
- 21 Plaintiffs’ Complaint, *Wagner v. Haslam*, No. 3:15-CV-115 at 2-3 (M.D. Tenn. 2015).
- 22 *Ibid.* at 22–23.
- 23 *Ibid.* at 22. The shift from a level five to one is an extreme shift. In those cases, the state or district should investigate to determine if there are large changes in student demographics or instructional staff that are affecting the measurement error. Depending on the findings, the state or district may wish to adjust their business rules for data inclusion.
- 24 *Ibid.*
- 25 The overall score includes the student achievement measures as well as the other criteria such as classroom observation. Under the original version of the policy, a teacher would need at least a level three school-wide growth score to obtain tenure, assuming he or she received a level five for the “other mandatory criteria” and that the school-wide growth is used as both the “student growth” and the “other measures of student achievement” component. Tennessee State Board of Education, Policy 5.201 (Apr. 19, 2013). *Wagner* at 23.
- 26 *Ibid.* at 24.
- 27 *Ibid.* at 25.
- 28 *Ibid.*
- 29 *Ibid.* at 26–27.
- 30 *Ibid.* at 26.
- 31 *Ibid.*
- 32 *Cook v. Stewart*, 28 F.Supp. 3d 1207 (N.D. Fl. 2014).
- 33 For instance, a physical education teacher could include mathematics concepts like geometry or science concepts such as physics. *Wagner* at 33. In the opinion, the court does speculate about the wisdom of encouraging teachers to incorporate subjects for which they are not licensed, but determines that is a policy decision for the legislature. Memorandum Motion to Dismiss. *Wagner* at 35.
- 34 The court offers that an effective physical education teacher may help students become healthier and more attentive in other classes and during exams. *Wagner* at 31–32. Or a good music teacher may help students build their mathematics skills or other subjects that require special focus. *Wagner* at 32.
- 35 *Ibid.* at 32.
- 36 *Ibid.* at 33.
- 37 *Cook v. Stewart*, 28 F.Supp. 3d 1207 (N.D. Fl. 2014).
- 38 The decision provides a few examples of what would be truly irrational. One included basing the Tennessee evaluation on if the Nashville Sounds minor league baseball team had a winning season. *Wagner* at 35.
- 39 *Ibid.* at 34.
- 40 *Ibid.* at 34.
- 41 *Ibid.* at 34.

42 Ibid. at 34.

43 Student portfolios require the collection of evidence throughout the school year. For example, in Tennessee, fine arts teachers are required to collect evidence such as student drawings or music performances at least four times during the school year. The collections must occur at at least two different points in time. They also must collect evidence of growth for both at the group level (e.g., a class) and a representation of students (i.e., emerging, proficient, and advanced). See Tennessee Department of Education, *Fine Arts*, [team-tn.org/non-tested-grades-subjects/fine-arts/](http://team-tn.org/non-tested-grades-subjects/fine-arts/).

44 The use of student learning objectives is “a participatory method of setting measurable goals, or objectives, based on the specific assignment or class, such as students taught, the subject matter taught, the baseline performance of the students, and the measureable gain in student performance during the course of instruction.” Race to the Top Technical Assistance Network, *Measuring Student Growth for Teachers in Non-Tested Grades and Subjects: A Primer* (Washington, DC: ICF International, 2010), [nassauboces.org/cms/lib5/NY18000988/Centricity/Domain/156/NTS\\_\\_PRIMER\\_FINAL.pdf](http://nassauboces.org/cms/lib5/NY18000988/Centricity/Domain/156/NTS__PRIMER_FINAL.pdf). Also

see Natalie Lacireno-Paquet, Claire Morgan, & Daniel Mello, *How States Use Student Learning Objectives in Teacher Evaluation Systems: A Review of State Websites* (Washington, DC: US Department of Education, 2014), [ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL\\_2014013.pdf](http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2014013.pdf).

45 See Croft & Buddin (2015).

46 Katie Buckley & Scott Marion, *A Survey of Approaches Used to Evaluate Educators in Non-Tested Grades and Subjects* (The Colorado Education Initiative, June 2, 2011).