

Teaching to the Test: Impacts of Evaluating Teachers Through Student Growth Measures

Dessi G. Kirova, Anton Scherbakov, Cynthia L. Blitz, & William A. Firestone

*Center for Effective School Practices, Rutgers University
New Brunswick, New Jersey, USA*

Funding received by states through the Race-to-the-Top program to implement systems for teacher evaluation is based on the requirement of using student growth measures as a significant portion evaluation ratings (U. S. Department of Education, 2009). In many states this has translated in basing about half of the teachers' professional evaluation ratings on student growth measures (Shepard, 2012). However, tying such large percentage of the teachers' evaluation to students' test scores, given that there is no evidence to support the assumption that teachers become better motivated to improve learning opportunities if evaluated on the basis of student test score gains, may have unintended consequences for teaching practices (Baker et al., 2010, p. 2).

As a result of the implementation of new systems for teacher evaluation, the pressure for teachers to better prepare students for standardized assessments has increased. Evidence shows that teachers shape their instruction to "conform to the expectations set by end-of-year accountability tests" (Shepard, 2012, p.15). Naturally, teachers react to the reality of being evaluated through their students' academic performance, and in this sense new evaluation systems have specific impacts on instructional practices. This paper explores and describes the actual changes in teachers' instructional practices as reported by teachers who have participated in a pilot teacher evaluation in a Northeastern state in 2012-2013 academic year.

The pilot teacher evaluation program was based on two different types of student academic measures: the

student growth percentiles (SGPs) and the student growth objectives (SGOs). The SGPs were introduced as a straightforward measure of student growth, calculated for each teacher of Math and/or English Language Arts (ELA) in grades 4-8. The SGPs measured individual student progress on statewide assessments (in at least two consecutive years) by tracking student scores from one year to the next to indicate how much the student changed relative to other students statewide with similar scores (2012-2013 Toolkit, 2012, p. 8). Alternatively, the SGO measures were based on teacher-selected growth objectives, and measuring student's progress along those objectives was done through teacher-selected or teacher-developed pre-assessments administered in the beginning of the school year and post-assessments, administered at the end of the academic year.

The differences in calculating student growth was a factor in teachers' perceptions of fairness, since SGP measures were to be used as a proportion of the evaluation for teachers of tested subjects (Math and ELA in grades 4-8) and SGO measures were to be used for all teaching staff in pilot-participating school districts. Additionally, the two measures accounted for a significantly different proportion of the final summative rating: SGPs were to account for 35%-45% of the evaluation rating, while the SGOs could account for as little as 10% of the rating.

Policy Impact on Instructional Practice

Previous studies have explored the impact of educational policy based on use of student achievement data on the instructional practices of teachers. Author 4 et al. (2004) found that when testing is used as a policy tool, it directly affects teachers' instruction. The study reported on data focused on the effect of the sanctions for teachers associated with use of student assessment data. Of interest is the study's discussion of factors contributing to increased pressure for teachers. The proposition that teachers' accountability systems, based on student

Access online from the Rutgers University Center for Effective School Practices. Suggested Citation:

Kirova, D. G., Shcherbakov, A., Blitz, C. L., & Firestone, W. A. (2015, April 17). *Teaching to the Test: Impacts of Evaluating Teachers Through Student Growth Measures*. 2015 Annual Meeting of the American Educational Research Association, Chicago, Illinois, USA. <http://tinyurl.com/kofuz56>

achievement have positive effect in that they motivate teachers to improve students' test results, is countered by evidence that these accountability systems have also detrimental effect on the actual quality of teaching, because of the emphasis on didactic (test-taking) instructional approaches, aimed at decontextualizing test preparation. This finding is also confirmed by evidence from research showing that decontextualized, test-like, drill-and-practice routines are harmful for students' learning (Shepard, 2012).

The situation is exacerbated in districts, where students' tests scores are consistently low. The sanctions for teachers based on student performance discourage teachers. In districts with large and unpredictable variation in student assessments, scores can undermine teacher morale. Evidence shows that "teacher attrition and demoralization have been associated with test-based accountability efforts, particularly in high-need schools" (Baker, et al., 2010, p. 4) Teaching in such districts adds increased pressure on teachers who must ensure increased students' tests performance scores to meet accountability standards (Firestone et al., 2004).

Based on our previous study (Firestone et al., 2004), we developed a broader theory about a range of factors that contribute to increased pressure for teachers evaluated partly on the basis of students' assessment scores. The next section offers a brief discussion of three factors contributing to added pressure: sense of controllability over the students' test scores, perceptions of purpose of the evaluation system (summative versus formative), and teaching in a low performing school districts.

Factors Contributing to Increased Teacher Pressure

In studying the teachers' fears regarding performance evaluation, Conley & Glasman (2008) concluded that teachers' doubts about whether they could control evaluated behaviors might inhibit their receptivity to evaluation. This in turn might raise the question of the impact of using both statewide assessments of student growth versus teacher-developed assessments, offering teachers entirely different levels of control over the test scores.

Secondly, the implementation of most teacher evaluation systems is motivated by both summative and formative propositions, asserting that teachers deserve

accurate evaluations which simultaneously provide opportunities for professional learning (2012-2013 Toolkit, 2012). However, assumptions that attaching direct consequences to students' test results affects instruction in that teachers become motivated to adopt more effective instructional practices do not necessarily hold true (Baker, et al., 2010; Shepard, 2012). Based on this and other relevant research (Author 4 et al., 2004), it is probable that teachers' perceptions of the purpose of the evaluation system, may contribute pressure on teachers to ensure students perform to their best ability on tests.

Lastly, multiple studies report that motivations for changing instructional practices for teachers being evaluated on the basis of students' scores and who teach in low-performing and under-resourced schools may be considerably different from teachers in better performing school districts (Baker, et al., 2010; Author 4 et al, 2004; Shepard, 2012). When faced with sanctions for students' low performance in historically low DFG-ranked schools (a state-specific indicator of districts' socioeconomic status), teachers are more likely to shift toward test preparation instructional practices.

Research Hypotheses

Based on the reviewed literature, this study is based on a broader understanding of factors that might be contributing to increased pressure, including teacher's perceptions of the purpose of teacher evaluation, understanding of controllability over students' test results and historical student performance data. Specifically the study explores the following hypotheses:

1. Teachers of tested subjects and grades would approach changing instruction under the new teacher evaluation system based on increased pressure for their students to score high on statewide tests.
 - a. Teachers' of tested subjects and grades perceptions of fairness of the evaluation systems would be less favorable.
2. Teachers who believe that the purpose of implementing systems for teacher evaluation is mainly to provide a summative rating (versus formative feedback) would adopt more didactic instruction.

- Teachers in historically low-performing school districts would be more likely to report adopting didactic instruction.

Method

The data for this study comes from an online survey of teachers from 25 school districts who participated in a teacher evaluation pilot program in 2012-2013 in a Northeastern state. The online survey was administered in June, 2013. The overall survey response rate was 39%.

The participants were 2,608 pre-K–12 teachers. The majority of respondents in the sample were female (77.6%) and most teachers surveyed had a Master's degree (54%, compared to 43% who had a Bachelor's degree). Importantly, a quarter of all survey participants were teachers of tested subjects.

Results and Discussion

To test the proposed hypotheses about the factors that might contribute to teachers' pressure and lead to teachers adopting more test-preparation methods to teach to their students, we constructed a Didactic scale, consisting of several items associated with test preparation, such as teaching strategies to score high on state or other standardized tests, emphasizing the importance of following procedures in solving mathematical problems or conducting scientific examinations, teaching the regular curriculum using items like those on the state tests or locally developed final exams, and motivating students to make their best efforts on the state tests or locally developed tests. When we conducted reliability analysis for the Didactic scale, alpha was .84.

To test the level of contribution of the hypothesized factors on didactic teaching, we conducted a multiple regression analysis by adding to the model the following predictors: 1. Teachers' role (teachers of tested versus untested subjects); 2. Perception of formative purpose; 3. Perception of summative purpose; 4. Perception of the possibility for losing tenure (for tenured teachers only); 5. District DFG factor as hypothesized factors contributing to pressure and to didactic approaches to teaching. The model has $R^2 = 0.87$, and $F(5, 1201) = 22.79$, $p < .01$. Regression coefficients for only three of

the predictor variables showed significance ($p < .01$): perceptions of formative purpose, being a teacher of tested subjects and DFG factor. The other two hypothesized predictors – perceptions of summative purpose and possibility of losing tenure did not significantly contribute to the model.

Based on these analyses we were able to confirm two of the proposed hypotheses. Specifically that teaching in low-performing districts contributes to added pressure and shifting toward didactic teaching, as teachers in low-performing school districts may feel under an increased pressure to change their teaching to ensure students score higher on assessments. Additionally, being a teacher of tested subjects significantly changes instructional practices toward more test preparation.

With respect to the effect of the teachers' perception of the main purpose of the teacher evaluation, we expected that teachers who believe that the main purpose of the evaluation is summative will be more likely to change their instruction toward test preparation. However, the perception of summative use turned out to be non-significant in predicting didactic teaching, in contrast to the perception of formative use. Additional correlation analyses revealed that both inquiry-oriented and didactic changes were significantly ($p < .01$) positively correlated to perceptions of formative use, and neither were significantly associated with perceptions of summative use. The fact that the teachers who believe that the purpose of evaluation is formative changed their practice to include both more test preparation and more inquiry-oriented teaching may be interpreted more generally. Teachers who reported "summative" belief may have understood the question as referring to implementation of teacher evaluation to comply with new policy by simply compiling evaluation ratings, instead of having to do with re-consideration of teaching practices, both didactic and inquiry-oriented.

Finally, we conducted t-tests for all fairness items after performing a Levene's test to reject the null hypothesis that variances were equal. Adjusting for the unequal variances, the difference between the two groups was significant such that than tested teachers believe that the SGO/SGP measures are more unfair than untested teachers, $t(1151) = -7.310$, $p < .01$. The t-tests showed

significance for all fairness items with $p < 0.01$, showing teachers of tested subjects were significantly less satisfied with the fairness of the way student achievement scores used to evaluate teaching effectiveness. We interpret this to be related to the teachers' sense of controllability over students' performance on assessments, and the fact that teachers of untested subjects self-selected or developed the assessments used to demonstrate students' growth.

Study Significance

The study offers an insight on the effects of implementing teacher evaluation systems, based on student achievement. The results from this study help understand better the motivations of teachers with respect to changes in teaching practice. The paper discusses some unintended consequences of the heightened accountability standards, specifically how increased pressures on teachers to prepare students for increased performance of assessments may change instruction and the quality of teaching. It also offers insight into the perceptions of fairness of evaluation ratings for teachers of tested and untested subjects.

References

- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*, EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.
- Blank, R. K. (2010). *State Growth Models for School Accountability: Progress on Development and Reporting Measures of Student Growth*. Council of Chief State School Officers.
- Conley, S., & Glasman, N. S. (2008). Fear, the school organization, and teacher evaluation. *Educational Policy*, 22(1), 63–85.
- Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: inquiry-oriented and didactic responses. *Assessment in Education*, 11(1), 67–88.
- Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems (Extended Version)*. Assessment and Accountability Comprehensive Center.
- National Council for Teacher Quality. (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Retrieved from: http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf
- New Jersey Department of Education. (2012). *2012–2013 Teacher Evaluation Pilot Toolkit*. Trenton, NJ: New Jersey Department of Education.
- Shepard, L. A. (2012). *Evaluating the Use of Tests to Measure Teacher Effectiveness: Validity as a Theory-of- Action Framework*.
- U. S. Department of Education. (2009, November). *Race to the top program executive summary*. Washington, DC.
- Whiteman, R. S., Shi, D., & Plucker, J. A. (2011). *Revamping the Teacher Evaluation Process*. *Education Policy Brief*. Volume 9, Number 4, Fall 2011. Center for Evaluation and Education Policy, Indiana University.



Graduate School of Education
Center For Effective School Practices

Teaching to the Test: Impact of Evaluating Teachers through Student Growth Measures

AERA 2015

Dessi Kirova
Cindy Blitz
Anton Shcherbakov
William Firestone

April 17, 2015

Abstract

Race-to-the-Top funding requires the use of student growth measures for teacher evaluation.

This study reports on specific changes in instructional practices, as reported by teachers.

Findings indicate several factors as significant contributors to instructional shifts toward increased test preparation:

- Use of SGP scores for evaluation purposes
- Being a teacher in a low-performing district
- Belief in the formative purpose of evaluations

2

The paper we will present today is focused on how teachers change their instruction as result of implementation of systems for teacher evaluation, partly based on student growth scores. As you know, under the RTT program, states who do not include student measures in teacher evaluation are likely to lose their federal funding.

We are interested in how teachers change their instruction under the pressure of preparing students to test well, for a lack of better word, and more specifically how different factors such as the types of assessments used to generate student growth scores and historic achievement in the school where they teach might impact the changes reported by teachers.

The study is based on the data from a pilot teacher evaluation system, implemented in 2011-2013.

Teacher Evaluation Pilot Program, 2012-2013

Student growth percentiles (SGPs)

SGPs measure individual student progress on statewide assessments (in at least two consecutive years) by tracking student scores from one year to the next to indicate how much the student changed relative to other students statewide with similar scores.

Student growth objectives (SGOs)

SGO measures are based on students' progress assessed through teacher-selected or teacher-developed pre-assessments (Fall) and post-assessments (Spring).

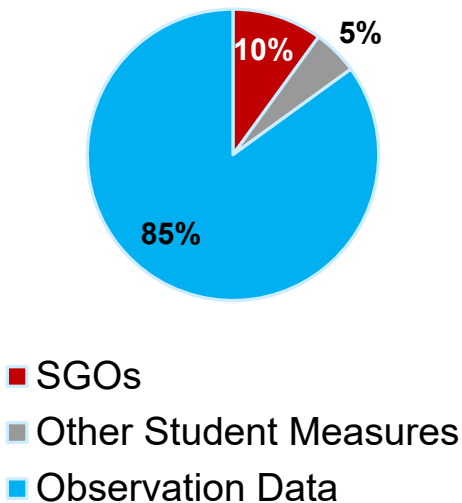
To situate the findings of the study, I'd like to give you some background about the pilot program, and more specifically about the measures used to account for student growth.

There were two types of measures through which student growth was measured. One – the SGPs were based on student performance on the state's standardized assessment, so it was used for all teachers of math and ELA in grades four to eight. The other one – SGOs were based on student performance on teacher-selected and teacher-developed assessments. These measures were mostly used for teachers on non-tested grades and subjects, so basically everyone who was not teaching math or language arts in fourth to eighth grade.

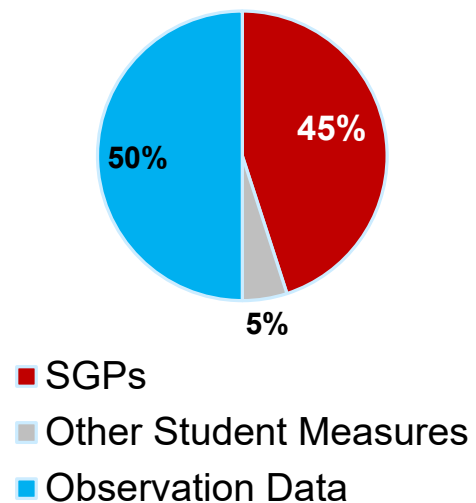
Teacher Evaluation Pilot Program, 2012-2013

Use of SGP versus SGO measures for evaluation of teachers of tested vs. non-tested subjects and grades.

Teachers of Non-tested Subjects & Grades



Teachers of Tested Subjects & Grades



4

Because the measures were not only different in terms of the assessment on which they were based, but were also used to account for a different proportion of the teachers summative ratings, we've included the pie charts on slide 4 to illustrate the extent of the potential difference in their weight.

SGP measures could account for as much as 45% of the summative score, whereas the SGOs could account for as little as 10%.

Literature Review

- Perceptions of control over aspects of the evaluation process might impact teachers' receptivity to evaluation (Conley & Glasman 2008).
- Teachers' perceptions of the summative purpose of the evaluation system may contribute to increased pressure to teach to the test (Firestone et al., 2004).
- Teachers in historically low-performing and under-resourced schools are more likely to shift toward instructional practices focused on test preparation (Baker, et al., 2010; Shepard, 2012).

5

To develop a better understanding of the factors that play a role in increasing the pressure for teachers to focus on test preparation, we have reviewed the literature to inform our thinking and expectations of the changes that might potentially result in more test preparation practices.

We were not surprised to find evidence that the perception of control over the aspects of the evaluation impacts teachers' receptivity to evaluation.

Additionally, we based our thinking on previous research of one of the authors of this paper, who has reported findings addressing the effect of beliefs about the summative purpose of teacher evaluation on increased pressure to prepare students for the test.

Finally, there is ample evidence in recent research suggesting that teachers in historically low performing schools are more likely to shift their instruction toward increased test preparation.

Research Hypotheses

1. Teachers of tested subjects and grades would change instruction under the new teacher evaluation system based on increased pressure for students to score high on statewide tests.
 - a) Teachers' of tested subjects and grades perceptions of fairness of the evaluation system would be less favorable.
2. Teachers who believe that the purpose of implementing systems for teacher evaluation is mainly to provide a summative rating (versus formative feedback) would focus more on test preparation.
3. Teachers in historically low-performing school districts would be more likely to report increased focus on test preparation.

6

Based on the review of the literature we constructed several research hypotheses with respect to how teachers might be changing their instruction as a result from the pilot teacher evaluation.

First, we hypothesized that teachers of tested subjects (the ones who are evaluated on the basis of their students performance on the state test) would be under more pressure to prepare students for testing. These teachers would also have a different perception of the fairness of use of student growth measures to assess teaching effectiveness.

We also hypothesized that perceived summative purpose would be associated with more test preparation.

And finally, teachers in historically low performing schools would be more likely to report increased focus on test preparation practices.

Method

Participants

- 2,608 preK-12 teachers
- A quarter of the sample were teachers of tested subjects

Data Source

- Online survey of teachers from 25 school districts, participating in a teacher evaluation pilot program
- Survey response rate was 39%

Moving to the method. One of the limitations of this study is that is based solely on teachers' self-reports. The data comes from a survey administered among all teachers participating in the pilot implementation. It was administered toward the end of the second year of pilot implementation in all 25 districts who took part in the pilot. The response rate was 39%.

It is important to note that about a quarter of the survey respondents were teachers of tested subjects.

Results and Discussion

Didactic Scale*

- Teaching testing strategies
- Importance of following procedures in solving mathematical problems or conducting scientific examinations
- Using items similar to state tests or locally developed final exams
- Motivating students to score high on state tests or locally developed tests

* Reliability analysis for the Didactic scale with $\alpha=.84$.

8

We began the analysis by constructing the so called Didactic scale. It includes response items to the question How much they have changed the use of a list of practices as a result of the teacher evaluation partly based on student growth measures. Some of the response items were:

- b. Teach strategies to score high on state or other standardized tests.
- c. Emphasize the importance of following procedures in solving mathematical problems or conducting scientific examinations.
- e. Teach the regular curriculum using items like those on the state tests or locally developed final exams.
- h. Motivate students to make their best efforts on the state tests or locally developed tests.

Results and Discussion

Multiple regression analysis with the following predictors:

1. Teachers' role (teachers of tested vs. non-tested subjects)
2. Perception of formative purpose
3. Perception of summative purpose
4. Perception of the possibility for losing tenure (for tenured teachers only)
5. District DFG factor* as hypothesized factors contributing to pressure and to didactic approaches to teaching

* State-specific indicator of districts' socioeconomic status.

9

We then ran a multiple regression analysis, using predictors as listed on slide 9:

- Teachers' role (teachers of tested vs. non-tested subjects)
- Perception of formative purpose
- Perception of summative purpose
- Perception of the possibility for losing tenure (for tenured teachers only)
- District DFG factor* as hypothesized factors contributing to pressure and to didactic approaches to teaching

Results and Discussion

The model has $R^2 = 0.87$, and $F(5, 1201) = 22.79$, $p < .01$.

Regression coefficients for three of the predictor variables were significant*:

1. Perceptions of formative purpose
2. Being a teacher of tested subjects
3. School district DFG factor

The other two hypothesized predictors – perceptions of summative purpose and possibility for losing tenure did not significantly contribute to the model.

* $p < 0.01$

10

We were able to confirm significant association between test preparation practices as exemplified by the Didactic scale, and three of our predictor variables:

1. Perceptions of formative purpose
2. Being a teacher of tested subjects
3. School district DFG factor

The other two hypothesized predictors perceptions of summative purpose and possibility of losing tenure did not significantly contribute to the model.

Formative purpose Q36. A major purpose of teacher evaluation in my district will be to provide information for teachers to improve their practice. (Strongly Agree to Strongly disagree)

Summative purpose Q37: A major purpose of teacher evaluation in my district will be to make tenure and promotion decisions. Q35. How do you think the new evaluation system will affect your chances of getting tenure?

- Greatly improve my chances, Improve my chances, Neither improve nor reduce my chances, Reduce my chances, Greatly reduce my chances, Don't know

Results and Discussion

- Perception of summative use was non-significant in predicting didactic teaching, in contrast to the perception of formative use.
- Additional correlation analyses revealed that both inquiry-oriented and didactic changes were significantly* positively correlated to perceptions of formative use, and neither were significantly associated with perceptions of summative use.

* $p < 0.01$

Results and Discussion

Based on conducted analyses we were able to support two of the proposed hypotheses:

1. Teaching in low-performing districts contributes to added pressure and didactic teaching.
2. Being a teacher of tested subjects significantly changes instructional practices toward more test preparation.

Results and Discussion

- We conducted t-tests for all fairness items.
- Teachers of tested subjects were significantly* less satisfied with the fairness of using student achievement scores to evaluate teaching effectiveness.
- We interpret this to be related to the teachers' sense of control over students' performance on assessments.

* $p < 0.01$

13

In terms of perceptions of fairness, we found that teachers of tested subjects were significantly less satisfied with the fairness of using student growth scores to assess quality of teaching. We think that this might have to do with the teachers' perceptions of control over the students' performance on the assessments used to calculate growth.

Study Significance

- Unintended consequences of the heightened accountability standards, such as how increased pressures on teachers to prepare students for enhanced performance of assessments may change the focus of instruction and the quality of teaching.
- Perceptions of fairness of evaluation ratings for teachers of tested and non-tested subjects are may also change the focus of instruction and the quality of teaching.

The study is significant in showing how new policy tying teaching effectiveness to student performance may change classroom instruction. Especially in tested subjects and grades and in historically low performing school districts.

References

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers, EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.

Blank, R. K. (2010). State Growth Models for School Accountability: Progress on Development and Reporting Measures of Student Growth. Council of Chief State School Officers.

Conley, S., & Glasman, N. S. (2008). Fear, the school organization, and teacher evaluation. *Educational Policy*, 22(1), 63–85.

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: inquiry-oriented and didactic responses. *Assessment in Education*, 11(1), 67-88.

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems (Extended Version). Assessment and Accountability Comprehensive Center.

National Council for Teacher Quality. (2011). State of the states: Trends and early lessons on teacher evaluation and effectiveness policies. Retrieved from: http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf

New Jersey Department of Education. (2012). 2012–2013 Teacher Evaluation Pilot Toolkit. Trenton, NJ: New Jersey Department of Education.

Shepard, L. A. (2012). Evaluating the Use of Tests to Measure Teacher Effectiveness: Validity as a Theory-of-Action Framework.

U. S. Department of Education. (2009, November). Race to the Top program executive summary. Washington, DC: Author.

Whiteman, R. S., Shi, D., & Plucker, J. A. (2011). Revamping the Teacher Evaluation Process. Education Policy Brief. Volume 9, Number 4, Fall 2011. Center for Evaluation and Education Policy, Indiana University.