Abstract Title Page
Not included in page count.

Title: *Tracking growth: Studying first-year teacher development under a high-stakes evaluation system*

Authors and Affiliations:

Adam Maier, Site Advisor-Analysis; The New Teacher Project
Erin Grogan, Partner, Assessment and Evaluation; The New Teacher Project

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Research demonstrates that the effectiveness of a student's teacher is the single most important school-based factor influencing achievement (Hanushek, Kain, & Rivkin, 2001; Sanders & Horn, 1994; Wright, Horn & Sanders, 1997; Gordon, Kane & Staiger, 2006). Students who have more effective teachers are more likely to attend college, earn a higher salary, and live in higher socioeconomic neighborhoods (Chetty, Friedman, & Rockoff, 2012). As such, teacher effectiveness is critically important, and identifying teachers who demonstrate high potential for growth in their first year of teaching could be a real asset to the districts in which they teach.

There is extensive evidence showing that teachers typically struggle in the early years of teaching when compared to their more experienced peers (e.g., Rivkin, Hanushek, & Kain, 2005; Kane, Rockoff & Staiger, 2006). However, prior research also shows that meaningful variation in performance exists (Nye, Konstantopolous, & Hedges, 2004; Odden, Borman, & Fermanich, 2004), leading to the reasonable assumption that not all first-year teachers struggle, and some show considerable potential for improving achievement of their students, both current (Gordon et al., 2006) and future (Rockoff & Speroni, 2009). The purpose of this project is to determine which teachers seem to measurably improve their instructional practice over the course of their first-year, measured via a series of observations conducted by normed observers using a common rubric; several studies (e.g., Jacob & Lefgren, 2008) have found observations to be predictive of student achievement.

The observation data are collected in the context of a high-stakes evaluation system designed to determine which teachers will be recommended for certification in their state at the end of their first year. Given that evaluations of teachers' instructional practices have been shown to significantly predict teachers' ability to promote student achievement (Bill & Melinda Gates Foundation, 2012; Kane et al., 2010; Sartain, Stoelinga, & Brown, 2011), and that assessments of instructional practice are linked to lasting improvements in performance even when attached to rewards and consequences (Taylor & Tyler, 2011), some improvement is expected. The more difficult problem in this study is attempting to untangle just which factors are significantly associated with higher levels of growth for some teachers, while others fail to improve.

Purpose / Objective / Research Question / Focus of Study:

This study sets out to answer the following research questions:

1. How reliably can growth be measured over the course of the first year of teaching?
2. Did teachers' instructions improve meaningfully?
3. Why did some teachers improve more than others during their first year of teaching?

Setting:

Data came from first-year teachers recruited and trained by alternative certification programs in 15 geographic regions: Delaware; Baltimore; Washington, DC; Chicago; Charlotte; Nashville; Memphis; Texas (Fort Worth, Dallas, Austin, and San Antonio); south Louisiana; and Georgia (Augusta, Savannah, and Southwestern districts).

All teachers were teachers of record in their classroom; that is, they were formally responsible for student learning despite their concurrent pursuit of certification while teaching in their first year. Teachers taught in all grades from kindergarten – grade 12, and taught in a

variety of content areas including math, English/language arts, physical sciences, special education, and foreign languages. Finally, teachers taught in both traditional public and charter schools.

Population / Participants / Subjects:

In total, 965 first-year teachers contributed data to the analysis. Of these teachers, about 41% taught elementary school, about 26% taught middle school, and about 33% taught high school. Subject areas were varied, with the largest group of teachers in a special education setting (21%), about 15% teaching secondary Sciences (Physical Science, Biology, Chemistry, and Physics), 14% teaching secondary math (primarily algebra and geometry), and about 10% teaching bilingual education or English as a second language. Fewer than 10% of teachers taught in each of the following subjects: Art; Early Childhood Education; Elementary Education; English/Language Arts; History/Social Studies; and Foreign Language (Japanese, French, etc.).

All teachers entered the classroom through alternative certification programs, with approximately 40% recruited through Teach for America, and the remaining 60% recruited through Teaching Fellows Programs.

Intervention / Program / Practice:

All teachers included in this study were participating in a one-year evaluation of classroom performance ultimately leading to a decision regarding whether or not to recommend the teacher for state certification. The evaluation included 3 classroom observations conducted by trained external observers; subjective performance evaluations provided by their principals; and, for teachers in certain grades and subjects, student survey data and value-added scores.

Classroom observations lasted 45-60 minutes. The first observation was announced, while the remaining sessions were unannounced. Observations were conducted using a rubric[1] which included seven teaching competencies each rated on a 5-point scale; ratings on each competency were then averaged, with performance falling into one of five categories[2]. Ratings on each of three observations were then averaged to arrive at a final observation score.

The 89 individuals who observed teachers were recruited and trained to use the rubric by the certification programs. Initial training involved a norming exercise by which observers had to demonstrate that they could meet an "80/60 threshold"[3] representing agreement with master coders. During the course of the observations, fidelity of using the rubric was checked periodically.

Research Design:

This study relies on a repeated-measures design, in which classroom performance data were collected on the same group of first-year teachers three times during the 2011-2012 school year. By collecting multiple measures of classroom performance for the same group of teachers over the course of the year, we are able to assess changes in performance while controlling for initial performance levels.

[1] See Appendix for observation rubric, including assessed teaching competencies.
[2] See Appendix for rating scale and corresponding performance categories.
[3] This meant that 80% of competency-level ratings for a given observation (7 of 9 competencies) were within one point of the master rating, and 60% of competency-level ratings (5 of 9 competencies) were an exact match with the master rating.

Data Collection and Analysis:

The primary data used in this analysis were collected via classroom observations. However, we supplement observation data with teacher perceptual data collected via surveys administered at the end of the school year.

Data were analyzed using a 2-level HLM:

Level 1: $$Y_{ti} = \pi_{0i} + \pi_{1i}(Observation\ Round)_{ti} + e_{ti}$$

Level 2: $$\pi_{0i} = \beta_{00} + r_{0i}$$
$$\pi_{1i} = \beta_{10} + r_{1i}$$

Where Y_{ti} is the observation score for person *i* in observation round *t*. The *Observation Round* variable was coded with the first observation as 0, the second observation as 1, and the third as 2, so that β_{00} represents the mean observation score in the initial observation and β_{10} represents the average amount of points gained in each successive observation. Of most interest to us in this analysis were the variances of the level-2 random effects, r_{0i} and r_{1i}, as these allowed us to explore how much variation existed in where teachers began their observation performance and how much they grew over time.

In some of the models we included predictor variables in the equations for π_{0i} and π_{1i} to see if we could account for where teachers began and how they grew over time, but the model above is the final version used to assess the variation in growth trajectories.

Findings / Results:

1. How reliably can growth be measured over the course of the first year of teaching?

- 32% of the variation in observation scores was due to persistent differences between teachers, suggesting observations in this evaluation system were as reliable as those in the MET study (Bill and Melinda Gates Foundation, 2012).
    - Improvement over time accounted for about 20% of the within-teacher variation.
- After assuming teachers will improve over time, 38% of the variation in observation scores is due to persistent differences between teachers.

2. Did teachers' instruction improve meaningfully?

- On average, teachers gained about 0.20 points with each additional observation, an increase that was statistically significant.
- However, teachers varied significantly in their growth trajectories, with some teachers actually showing a negative trajectory over the course of the year.
- A teacher with a growth trajectory 1 SD above average was predicted to have a 3rd observation score about 0.5 units higher than a teacher with a growth trajectory that was 1 SD below average.

3. Why did some teachers improve more than others during their first year of teaching?

- Teachers' initial observation score was correlated with their rate of growth.
    - Among all teachers, one's initial observation score is negatively correlated with growth. ($r \approx -0.061$)
        - When analysis is restricted to teachers with initial observation scores in the middle 80%, the association between initial score and growth is positive and weaker ($r \approx 0.038$)

- One particular rubric indicator-*Facilitates organized, student-centered, objective-driven lessons*- seemed to be an important signal of teacher growth.
    - Teachers who earned a 4 or 5 on this competency in their first observation improved at a rate of 0.18 points per observation – a rate significantly greater than those who initially earned a 3 on this indicator.
    - However, teachers who initially earned a 1 or 2 on this indicator, tended to worsen by 0.10 points per observation.
- There were minor differences in growth by subject, with math teachers improving the least.
- There were significant differences in teacher growth based on geographic area.
    - In general, growth ranged from 0.12 – 0.24 across geographic locations.
    - If teachers were randomly assigned to sites, this range would be about 0.19-0.21.
- Some teacher perceptions (satisfaction with school; self-efficacy) were significant predictors of *initial* observation scores, but were not associated with growth.
- There were no differences in growth between teachers entering the classroom through different certification programs, nor were there differences based on perceived fairness of the evaluation system.

Conclusions:

This sample of almost 1,000 geographically diverse first-year teachers, all evaluated by trained observers using the same rubric, showed evidence that teachers vary meaningfully in their classroom performance in their first year of teaching. Early proficiency was the best predictor of overall performance, suggesting that interventions could be targeted early in the year to support teachers who are struggling from the beginning of assuming responsibility for their own classroom. Teachers who seemed adept at facilitating student centered-lessons appeared able to achieve more growth in performance than their peers who struggled at this competency. Further, math teachers seemed to perform less well than teachers in other subjects, suggesting that either a) the rubric (or observers) is not well equipped to assess math performance or b) math teachers recruited and trained by these certification programs are not receiving the necessary supports to perform well in the classroom.

However, our data allow us to explain only about 38% of the between-teacher variation, leading to concerns about omitted variables and their effects on our outcomes. Further, our lack of understanding regarding which factors are associated with so much of the variation between teachers has implications for the certification programs training and supporting these teachers; specifically, our dataset does not include critical information regarding the specific interventions undertaken with individual teachers. These programs are urgently interested in targeting teachers most at risk for low performance as early in the year as possible, but have only limited information on which to act. With the exception of math teachers, there was little variation in growth between teachers of different subject areas, which is not particularly helpful for targeting content-specific interventions. Further, teacher perceptions of their own teaching efficacy and their satisfaction with their teaching environment showed no association with growth, making it difficult to target teachers who might be at risk of struggling by asking them directly about their teaching experience.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bill and Melinda Gates Foundation (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. MET project, Research Paper. Retrieved March 30, 2012 from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Chetty, R., Friedman, J., & Rockoff, J. (2012). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. *NBER Working Paper 17699*. Cambridge, MA: National Bureau of Economic Research.

Gordon, R., Kane, T., & Staiger, D. (2006). Identifying effective teachers using performance on the job. *Brookings Institute Discussion Paper 2006-01*.

Hanushek, E., Kain, J., & Rivkin, S. (2001). Why public schools lose teachers. *NBER Working Paper 8599*. Cambridge, MA: National Bureau of Economic Research.

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

Kane, T. J., J. E. Rockoff and D. O. Staiger (2006). What does certification tell us about teacher effectiveness? Evidence from New York City. *NBER Working Paper 12155*. Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., Taylor, E. S. Tyler, J. H., & Wooten, A. L. (2010). Identifying effective classroom practices using student achievement data. *NBER Working Paper 15803*. Cambridge, MA: National Bureau of Economic Research.

Nye, B., Konstantopoulos, S, & Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32.

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

Rockoff, J., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(2): 261-266.

Sartain, L., Stoelinga, S., & Brown, E.R. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. Chicago, IL: Consortium on Chicago School Research. Retrieved March

30, 2012 from http://ccsr.uchicago.edu/publications/Teacher%20Eval%20Report%20FINAL.pdf

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299–311.

Taylor, E. S., & Tyler J. H. (2011). The effect of evaluation on performance: evidence from longitudinal student achievement data of mid-career teachers. *NBER Working Paper 16877*. Cambridge, MA: National Bureau of Economic Research.

Wright, P. S., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.

Appendix B. Tables and Figures
Not included in page count.

Figure 1. Sample Observation Report

ASSESSMENT OF CLASSROOM EFFECTIVENESS

Observation Score Report

Below you will find your competency scores from the classroom observation conducted by your ACE Observer this past week. After visiting your classroom, the ACE Observer assigned a numerical rating for each competency based on the teaching practices he or she observed during this lesson. Please refer to your copy of the ACE Framework and the accompanying rubric for a more detailed understanding of what each competency addresses.

SCORING

(1) = Ineffective; (2) = Minimally Effective; (3) = Developing; (4) = Proficient; (5) = Skillful

OBSERVATION INFORMATION

Participant Name	John Smith
Observer Name	Daffy Duck
Date of Observation	1/31/2012
Observation # (1, 2, 3)	1
Site/Program	TNTP Academy
School	P.S. 123

ACE FRAMEWORK COMPETENCY RATINGS

COMPETENCY	SCORE
I-1: Facilitates organized, student-centered, objective driven lessons	3
I-2: Checks for student understanding and responds to student misunderstanding	3
I-3: Differentiates instruction by employing a variety of instructional strategies	2
I-4: Engages student sin work that develops higher-level thinking skills	2
I-5: Maximizes instructional time	3
I-6: Communicates content and concepts to students	2
I-7: Promotes active participation and high academic expectations	3
E-1: Sets and implements discipline management procedures	2
E-2: Builds a positive and respectful classroom environment	4
OVERALL SCORE FOR THIS OBSERVATION:	2.67

QUESTIONS ABOUT ACE OBSERVATIONS

If you have any questions about your ACE observation score or ACE observations in general, please contact Arne Duncan (Site Manager) at arne.duncan@ed.gov.

A NOTE ON YOUR OBSERVATION SCORE

- Through ACE, x Teaching Fellows (XTF) strives to create the fullest possible picture of your performance and to help you be more successful by giving you insights into your practice.
- You must achieve an overall score at the Developing level or higher to pass the observational component of ACE, and you can expect to be observed at least 2-3 times this year.
- It is important to understand that all of your observation scores will be averaged together to calculate your overall ACE observation score. In turn, this final observation score will be considered alongside other measures such as your principal rating to calculate your final ACE rating.
- This report provides a valuable opportunity for you to reflect on your strengths and weaknesses and identify steps you can take to improve your teaching, particularly in any competency areas where you scored a 2 or lower. We encourage you to revisit the ACE Framework to get a better understanding of what performance at the Developing level looks like and consider how to incorporate those behaviors and practices into your classroom.

If you have further questions about resources available to help you improve your instructional practice, please contact Arne Duncan (Site Manager) at arne.duncan@ed.gov.

Table 1
Performance categories and associated point ranges.

Ineffective	1.00-1.99
Minimally Effective	2.00-2.79
Developing	2.80-3.59
Proficient	3.60-4.29
Skillful	4.30-5.00