What's Happening

January 2016

Analysis of the stability of teacher-level growth scores from the student growth percentile model

Andrea Lash
Reino Makkonen
Loan Tran
Min Huang
WestEd

Key findings

Some states that evaluate teachers based partly on student learning use the student growth percentile model, which computes a score that is assumed to reflect a teacher's current and future effectiveness. This study in a Nevada school district finds that half or more of the variance in teacher scores from the model is due to random or otherwise unstable sources rather than to reliable information that could predict future performance. Even when derived by averaging several years of teacher scores, effectiveness estimates are unlikely to provide a level of reliability desired in scores used for high-stakes decisions, such as tenure or dismissal. Thus, states may want to be cautious in using student growth percentile scores for teacher evaluation.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL WEST
Regional Educational Laboratory
At WestEd

U.S. Department of Education
John B. King, Jr., *Acting Secretary*

Institute of Education Sciences
Ruth Neild, *Deputy Director for Policy and Research*
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Joy Lesnick, *Acting Commissioner*
Amy Johnson, *Action Editor*
OK-Choon Park, *Project Officer*

REL 2016–104

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

January 2016

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0002 by Regional Educational Laboratory (REL) West at WestEd. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Lash, A., Makkonen, R., Tran, L., & Huang, M. (2016). *Analysis of the stability of teacher-level growth scores from the student growth percentile model* (REL 2016–104). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from: http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

States across the nation are developing new systems to evaluate teachers based on classroom observations, how much students learn, or some combination of these and other factors. Evaluations under these systems can have high stakes for teachers. Poor evaluations may lead to a frozen salary, mandatory remediation, or dismissal, while exceptional evaluations may be rewarded with a salary increase or tenure.

This study tests an implicit assumption of high-stakes teacher evaluation systems that use student learning to measure teacher effectiveness: that the learning of a teacher's students in one year will predict the learning of the teacher's future students. Evaluation systems that identify low-scoring teachers for remediation assume that if the teachers are not retrained, their future teaching will also be relatively ineffective. Systems that award tenure to teachers who score higher assume that those teachers will remain effective. Examining the stability of teacher-level growth scores over time provides evidence of the validity of the interpretation and use of such scores for teacher evaluation and offers information that could be useful in designing alternative evaluation systems.

A common method of measuring student learning for teacher evaluation is the student growth percentile model, which assigns each student a percentile rank in the distribution of assessment scores for students at the same grade level and with a similar achievement history. The median student growth percentile of a teacher's students is the teacher-level growth score, which tends to be used for teacher evaluation.

This study, requested by the Nevada Department of Education, investigates the stability of the teacher-level growth score. Three years of math and reading score data were analyzed for close to 370 elementary and middle school teachers from Washoe County School District, Nevada's second largest school district.

In math, half the variance in teacher scores in any given year was attributable to differences among teachers, and half was random or unstable. In reading, the proportion of the variance attributable to differences among teachers was .41, and .59 was random or unstable.

More stable measures of effectiveness can be constructed by averaging multiple years of growth scores for a teacher. For example, when effectiveness is computed as an average of annual scores for three years, the proportion of the variance in teacher scores attributable to differences among teachers is .75 in math and .68 in reading.

These estimates do not meet the .85 level of reliability traditionally desired in scores used for high-stakes decisions about individuals (Haertel, 2013; Wasserman & Bracken, 2003). States that are considering the student growth percentile model for teacher accountability may want to be cautious about using the scores for high-stakes decisions.

Contents

Why this study?

As of early 2014, 40 states and the District of Columbia were using or piloting methods to evaluate teachers in part according to the amount students learn (Collins & Amrein-Beardsley, 2014). Such evaluations can have high stakes for teachers. In some states the consequences of a poor evaluation can include a frozen salary, remediation, or dismissal, while consequences for exceptional evaluations can include a bonus, a salary increase, or tenure (Herlihy et al., 2014).

A common method of measuring student learning for teacher evaluation is the student growth percentile model developed by Betebenner (2011), which is sometimes referred to as the Colorado growth model. It is in various stages of use—from preliminary investigation to full-scale adoption—in as many as 20 states (New Jersey Department of Education, 2012; see also Collins & Amrein-Beardsley, 2014[1]). This study—requested by the Nevada Department of Education, which at the time of writing was planning to use the student growth percentile model—investigated the stability over time of teacher-level growth scores, which are derived from student growth scores under the student growth percentile model. While other research has examined the stability of both school-level scores derived from the student growth percentile model and teacher effectiveness measures derived from value-added models (see appendix A), the study team could not locate any published studies on the stability of teacher-level growth scores derived from the student growth percentile model.

Examining the stability of teacher-level growth scores offers information that could be useful in selecting, weighting, and combining measures in evaluation systems

The stability of teacher-level growth scores is important to evaluation systems that use the scores to measure teacher effectiveness. Underlying such systems is the implicit assumption that a teacher's growth score in one year predicts that teacher's effectiveness in future years (Glazerman et al., 2011). Evaluation systems that identify low-scoring teachers for remediation assume that if the teachers are not retrained, their future scores will remain low. Similarly, systems that award tenure to teachers who score higher assume that those teachers will continue to be effective. Examining the stability of teacher-level growth scores provides evidence of the extent to which this assumption is warranted and offers information that could be useful in selecting, weighting, and combining measures in evaluation systems.[2]

In Nevada, teacher-level growth scores are included in accountability models. Initially used only for schools, they are now used for teachers as well. In 2009 the Nevada Legislature mandated a statewide growth model for school accountability. The Nevada Department of Education established selection criteria for the model, including that it be a valid, reliable, and technically sound metric conditioned on students' past performance when performance is measured by scores that are not on the same scale from one grade to the next (Davidson, Ozdemir, & Harris, 2010). The department ultimately selected the student growth percentile model. In this model, growth is not measured as the change in a student's test scores from one year to the next but as the percentile rank of the student's score in the distribution of the current year's achievement scores for all students in the state who are at the same grade level and who have similar past performance. A grade 6 student's growth score of 40 for math in 2010 would indicate that the student had a 2010 math achievement test score equal to or higher than those of 40 percent of the state's grade 6 students who had a math achievement history similar to the student's.[3] A student growth score of 50 (the median of the distribution) would indicate typical growth, with higher or

lower scores indicating greater or less than typical growth (Nevada Department of Education, 2010).

In 2011 the Nevada State Legislature expanded the use of growth-model data to educator evaluation. Student growth scores are used to produce a teacher-level growth score, which is the median of the growth scores for the teacher's students. A score below 50 indicates that the teacher's typical student scored lower than would be expected for students who started the year at a similar achievement level.

Like other states, Nevada is planning to include in its teacher evaluation system multiple measures of teacher effectiveness and multiple years of student outcome data. The analysis of the stability of teacher-level growth scores in this study can inform decisions in Nevada and elsewhere about how to incorporate teacher-level growth scores as a measure of student learning.[4]

Analysis of the stability of teacher-level growth scores can inform decisions about how to incorporate teacher-level growth scores as a measure of student learning into teacher evaluation systems

What the study examined

This study examines one overarching research question: How stable over years are annual teacher-level growth scores, derived by applying the student growth percentile model to student scores from Nevada's Criterion-Referenced Tests in math and reading? In other words, how likely is it that the same score would be obtained in different years?

The data, methods, and summary variables used in the report are discussed in box 1. Details on how the student achievement variables were constructed are in appendix B, and details on the design and methods are in appendix C.

Box 1. Data, methods, and summary variables

Data

Because no statewide datasets were available that linked students to teachers so that teacher-level scores could be computed, the study uses data on all students in grades 4–8 from Washoe County School District—Nevada's second largest school district, with a student enrollment of more than 60,000. The district provided student-level scores linked to teachers for three school years beginning in 2009/10. Data included the student's grade, school, school level (elementary or middle), teacher (math or English language arts), class (because some teachers teach more than one class and some classes are taught by more than one teacher), current year's score on Nevada's Criterion-Referenced Test, proficiency level associated with that score, growth score (that is, student growth percentile) for the current year, and an indicator variable that identifies whether the student was enrolled in the school for the full school year. A teacher-level dataset for each school year was then prepared that contained the teacher ID, school ID, and four variables derived from the student achievement data: the percentage of the teacher's students who were proficient in reading, the percentage who were proficient in math, the teacher-level growth score in reading, and the teacher-level growth score in math (see appendix B for more information on the variables created for the study).

(continued)

Box 1. Data, methods, and summary variables *(continued)*

Methods

The analysis is based on a generalizability study, which partitions the variance in teacher scores into components and estimates the magnitude of each (Brennan, 2001; Shavelson & Webb, 1991; see appendix C). Here the three relevant components of variance are true differences among teachers, which may provide useful information for decisionmakers; systematic year-to-year fluctuations, which affect all teachers' scores; and random and other sources of instability, which cause teachers' scores to change in unsystematic ways from year to year. The last two components could lead to errors in evaluating teachers.

Summary variables

The stability of the teacher scores over time is summarized using the reliability coefficient, which is the proportion of the total variance in scores that is attributed to the first component of variance, the true differences among teachers.[1] The coefficient is reported on a scale of 0 to 1, where 0 means that none of the variance is due to true differences between teachers and 1 means that all the variance is due to true differences. The higher the reliability coefficient, the more stable the scores. For high-stakes decisions about individuals, some researchers argue for a reliability coefficient of .85 or higher (Haertel, 2013; Wasserman & Bracken, 2003). By comparison, scores for the licensing examination required for nurses are estimated to have reliability coefficients of .87–.92 (National Council of State Boards of Nursing, n.d.), and scores for college admissions tests have reliability coefficients ranging from .89 to .93 (College Board, 2013).

The magnitude of the total error in scores—errors that are associated with the second and third components of variance—is summarized using the standard error of measurement. It can be used to establish a score range that is likely to represent a teacher's effectiveness, much like a margin of error is established for results from opinion polls. One commonly used margin of error is the 95 percent confidence interval—that is, a range of scores within which there is a 95 percent chance that the true score lies (Salkind, 2008). The upper end of this range is obtained by adding 1.96 times the standard error of measurement value to the observed score, while the lower end is attained by subtracting 1.96 times the standard error of measurement from the observed score. Because the standard error of measurement is reported in the units of the score scale that is being evaluated, it cannot be compared across measures with different scales, which means that, unlike for reliability coefficients, there are no general guidelines for a target standard error of measurement or 95 percent confidence interval.

The study also reviews the stability of the status score—the proportion of a teacher's students who meet grade-level standards—which was used in Nevada's accountability systems before growth scores were introduced. The status score is more familiar to many educators and is included in some states' educator evaluation models, so comparing its stability to that of the teacher-level growth score is of some interest.

Note

1. Depending on the methods used to estimate this proportion, the coefficient might be referred to as a reliability coefficient, a generalizability coefficient, or a stability coefficient. These distinctions are not important to this report, as the interpretation would be the same no matter how the coefficient is labeled. To simplify the presentation, this report uses the general term reliability coefficient. The calculations are for an absolute comparison and not a relative comparison, as explained in appendix C.

What the study found

Nevada's annual teacher-level growth scores, derived by applying the student growth percentile model to student scores from Nevada's Criterion-Referenced Tests in math and reading, did not meet a level of stability that would traditionally be desired in scores used for high-stakes decisions about individuals.

This section presents the basic results from the analysis of the stability of annual teacher-level growth scores. The statistics underlying the findings are then used to extrapolate the likely results if multiple years of data were averaged to derive a teacher score. Additional extrapolations project the likely misclassification rates for a particular cutscore; the misclassification rates provide a practical example of the consequences of using teacher-level growth scores with low stability.

Half or more of the variance in teacher-level growth scores was due to random or otherwise unstable fluctuations

The proportion of the variance among teachers in any given year that was attributable to true differences among teachers was .50 in math and .41 in reading; the rest was due to random or otherwise unstable fluctuations

No more than half the variance in annual teacher-level growth scores was due to true differences among teachers. The proportion of the variance in any given year that was attributable to true differences among teachers was .50 in math and .41 in reading; the rest was due to random or otherwise unstable fluctuations (table 1).

The reliability coefficient for status scores was .64 for math and .65 for reading (see table 1). Thus, the proportion of the variance that was random or otherwise unstable was .36 for math and .35 for reading.

The range of scores likely to include a teacher's true score would span close to half the 100 point scale

For the annual teacher-level growth scores, the standard error of measurement was 12.22 for math and 11.31 for reading (see table 1). This means that the 95 percent confidence interval for a teacher's true score would span 48 points for math, a margin of error that covers nearly half the 100 point score scale, and 44 points for reading. For example, one would be 95 percent confident that the true math score of a teacher who received a score of 50 falls between 26 and 74. More precision would be obtained with measures that are more stable.

Even when derived by averaging three years of teacher scores, effectiveness estimates based on student growth are unlikely to provide a level of stability desired for use in high-stakes decisions

The stability of a score increases when the score is derived from averages taken over two or three years of data (see table 1).[5] This is because the positive and negative errors in annual growth scores are averaged, reducing their effect on the teacher-level growth score. For example, if the teacher-level growth score were computed using an average of three years of data (the maximum number of years before tenure is determined in Nevada), the coefficients would be .75 for math and .68 for reading—higher than the coefficients that were found for a single annual score (figure 1).

Table 1. Variance components, coefficients, and standard errors of measurement for teacher-level growth scores and teacher-level status scores derived from student achievement scores in math and reading

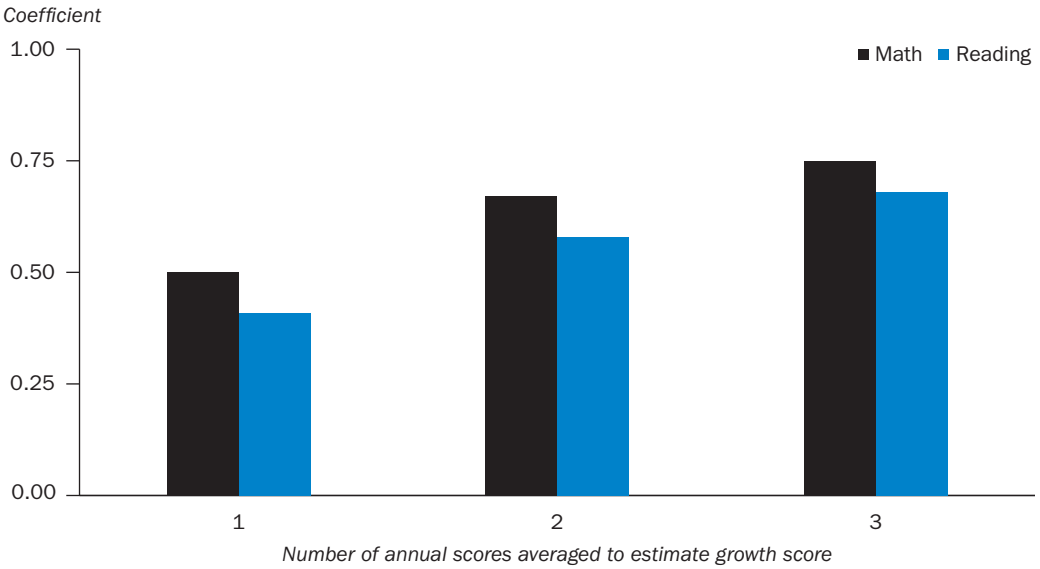
Estimates derived from generalizability study	Teacher-level growth score (median student growth percentile)		Teacher-level status score (percentage of students who meet grade-level standards)	
	Math (n = 369)	Reading (n = 375)	Math (n = 369)	Reading (n = 375)
Variance component				
Teacher	151.71	90.51	210.40	260.58
Year	8.87	3.57	8.16	1.50
Residual	140.50	124.41	108.27	141.10
Total	301.08	218.49	326.83	403.18
Coefficient[a]				
Score from one year	.50	.41	.64	.65
Average of scores from two years	.67	.58	.78	.78
Average of scores from three years	.75	.68	.84	.85
Standard error of measurement[b]				
Score from one year	12.22	11.31	10.79	11.94
Average of scores from two years	8.64	8.00	7.63	8.44
Average of scores from three years	7.06	6.53	6.22	6.89

a. Calculated as teacher component/[(teacher component) + (year component/k) + (residual component/k)], where k is the number of annual teacher-level scores that are averaged to produce the final teacher score.

b. Calculated as the square root of [(year component/k) + (residual component/k)], where k is the number of annual teacher-level scores that are averaged to produce the final teacher score.

Source: Authors' analysis of 2009/10, 2010/11, and 2011/12 data provided by Washoe County School District.

Figure 1. The stability of teacher-level growth scores increases when more annual scores are averaged



Source: Authors' analysis of growth scores (from 2009/10, 2010/11, and 2011/12 data) provided by Washoe County School District.

About one in five "typical" math teachers would be expected to be misclassified as low performing when ineffective teaching is defined as an annual growth score less than 40

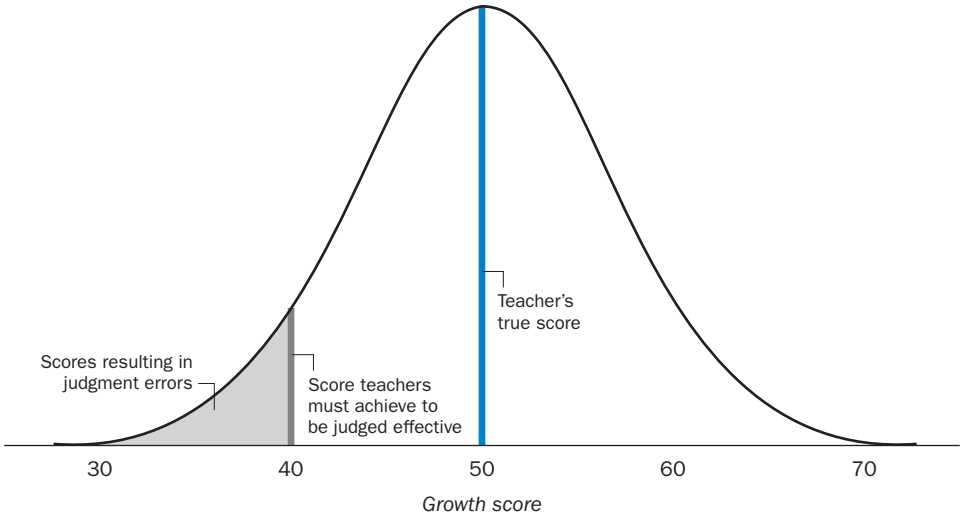
A teacher with a growth score of 50 is "typical" in the sense that the teacher's median student had a growth score at the median of students in the state with a similar achievement history. The teacher's growth scores as measured on any one occasion may differ from 50 due to measurement error. Considering the possible scores a teacher could receive on a large number of different occasions, the proportion of the teacher's scores that are below a specified cutscore for effectiveness, say 40,[6] is the proportion of times the typical teacher would be misclassified as ineffective (figure 2).

It is convenient and common to assume a normal distribution of these scores that might be observed for a given teacher (Harvill, 1991). In a normal distribution with a mean of 50 (the teacher's true score) and a standard deviation of 12.22 (the standard error of measurement for the annual teacher-level growth score in math, as noted above), 21 percent of the scores in the distribution fall below 40; thus, the typical teacher would be expected to be misclassified as ineffective 21 percent of the time, when classifications are based on a single year of student growth.

Assuming a normal distribution of scores and a cutscore for effectiveness of 40, a teacher with a growth score of 50 would be expected to be misclassified as ineffective 21 percent of the time, when classifications are based on a single year of student growth

The same potential for misclassification applies when talking about a teacher whose true score falls below the cutscore for effectiveness but who, due to measurement error, is identified as effective. For more on how the results of a reliability study can be used to estimate the number of teachers who are likely to be misclassified in decisions about their effectiveness, see appendix D.

Figure 2. Hypothetical distribution of possible growth scores for a typical teacher with true growth score at 50



Source: Authors' construction.

Implications of the study

This is perhaps the first published study of the stability of the teacher-level growth score derived under the student growth percentile model, a common model used by states in teacher evaluation systems. States or districts may have conducted studies to explore the model; if so, those studies have not appeared in the published literature or been posted on the Internet (see appendix A for related research). The findings indicate that even when computed as an average of annual teacher-level growth scores over three years, estimates of teacher effectiveness do not meet the level of stability that some argue is needed for high-stakes decisions about individuals, which is a coefficient of .85 or higher (Haertel, 2013; Wasserman & Bracken, 2003). The current finding that teacher-level growth scores are so unstable as to raise questions about their use in teacher evaluation systems is similar to conclusions that other researchers have drawn about value-added measures of teacher effectiveness (for example, American Educational Research Association, 2015; American Statistical Association, 2014; Haertel, 2013; Konstantopoulos, 2014). And the finding is consistent with research about classroom teaching that has documented how teachers' effects on student learning vary over many dimensions of the classroom, including subject matter, students, and occasions (Berliner, 2014; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011; Good, 2014).

As states examine properties of their estimates of teacher effectiveness and decisionmakers weigh how to incorporate teacher-level growth scores in teacher accountability policy, they may want to exercise caution and further investigate whether teacher-level growth scores are sufficiently stable for use in high-stakes decisions

The conclusion that growth scores alone may not be sufficiently stable to support high-stakes decisions suggests the need to examine measures of teacher effectiveness and their interpretation in evaluation systems. The growth score may not be a sound measure of a teacher's effectiveness, or the magnitude of a teacher's effect on student learning may not be as predictable a trait of the teacher as many evaluation systems assume it is. Rather, a teacher's effectiveness may depend in part on features of the teacher's students—that is, the collection of students in any given year, which change from one year to the next (for example, Guarino, Reckase, Stacy, & Wooldridge, 2014). Growth measures may need to be thought of differently—considered a measure that is associated with a particular combination of teacher and students rather than one that is attributable to the teacher alone (E. Haertel, personal communication, 2012). Thus, as states examine properties of their estimates of teacher effectiveness and decisionmakers weigh how to incorporate teacher-level growth scores in teacher accountability policy, they may want to exercise caution and further investigate whether teacher-level growth scores are sufficiently stable for use in high-stakes decisions.

Many educator evaluation models include multiple measures such as teacher observations, surveys, or additional student outcomes. So policymakers may want to consider the stability of those other measures and examine the reliability of different combinations of measures and the weight assigned to different measures. The methods used in this study to extrapolate the stability of different numbers of years of data or misclassification rates may be of interest to policymakers as they consider how to refine their educator evaluation models. Local reliability statistics can be used in ways analogous to those illustrated here to test the reliability of different scenarios under consideration.

Limitations of the study

This study has three main limitations.

First, the study examines scores from teachers in one Nevada school district rather than from the statewide population of Nevada teachers. This is because no statewide dataset is available that links teachers to students so that teacher scores can be derived from student-level data. Examining a single district may affect the study in two ways:

- If Washoe County School District teachers are more homogeneous in teacher effectiveness than the population of Nevada teachers, the study would underestimate the stability of the scores. Similarly, if Washoe County School District teachers are less homogeneous in teacher effectiveness than the population of Nevada teachers, the study would overestimate stability.
- If Washoe County School District teachers' scores are more volatile, on average, from year to year than those of other teachers in the state, the study would underestimate stability. Similarly, if Washoe County School District teachers' scores are less volatile, on average, from year to year, the study would overestimate stability.

Second, to examine stability of scores, the study used teachers who have multiple years of scores. The sample of teachers who remained in the district teaching a particular topic (reading or math) in grades 4–8 for the period of study may differ from a sample with teachers who changed districts or moved to untested grades or subjects. For example, teachers who stay in the same setting may have the advantage of adapting to that setting and may have more stable scores.

Third, the study examines scores from Nevada's Criterion Referenced Tests, which were the state's assessment for accountability purposes until 2015/16. Like other states, Nevada is moving to new assessments based on the Common Core Standards. How that change will affect teacher-level growth scores and the stability of those scores is unknown.

Appendix A. Related literature

Two bodies of research are relevant to this study. One is prior studies of the stability of the student growth percentile model, the model that is the focus of this study. Those studies focus on the stability of school-level scores; they have not examined teacher-level scores. The second is studies of the stability of teacher effectiveness as measured through value-added models. While value-added models take a different analytic approach from the student growth percentile model, they also derive a teacher effectiveness score, and are used in some states and districts as part of their educator effectiveness models.

Prior to this study, no research had been published on the stability of teacher-level growth scores derived from the student growth percentile model. But related studies have examined the stability of both school-level student growth percentile scores and teacher-effectiveness measures derived from value-added models. Goldschmidt, Choi, and Beaudoin (2012) estimated the year-to-year stability in school-level growth scores derived from the student growth percentile model by correlating school scores from two consecutive years. They found correlations of .46 for math in both elementary and middle school samples. The correlations for reading were lower: .32 for elementary schools and .22 for middle schools. Lash, Peterson, Vineyard, Barrat, and Tran's (2013) recent generalizability study found similar results for school-level student growth percentile scores. They analyzed four years of growth scores for the population of elementary and middle schools in Nevada and found results (comparable to the correlations in the Goldschmidt et al. [2012] study) of .43 for math and .38 for reading. They examined the implications of these results for the accuracy of decisions in a school accountability system designed to identify low-performing schools and found that if schools with annual school-level growth scores below 40 were classified as low-performing schools, 14 percent of the classifications in math and 11 percent in reading would likely have been in error.[7] In other words, these percentages of schools were likely to have been misclassified solely because of measurement instability.

Other research has examined the year-to-year stability of estimates of teacher effects derived from value-added models, a popular alternative to the student growth percentile model. In analyzing data from five of Florida's largest school districts, McCaffrey, Sass, Lockwood, and Mihaly (2009) found considerable year-to-year variance in teachers' value-added estimates, even after accounting for some factors that could change annually, such as experience and recent in-service professional development. Roughly a third of the top 20 percent of Florida teachers remained in the top 20 percent the next year, while approximately a tenth of those who had been in the top 20 percent fell to the bottom 20 percent of the teacher effectiveness distribution the next year.

Other recent studies have shown similar year-to-year shifts in teachers' value-added rankings. Newton, Darling-Hammond, Haertel, and Thomas (2010) found that 19–41 percent (depending on the value-added model used) of teachers saw their effectiveness rankings shift by three or more deciles (that is, by 30 percent or more of the population) in either direction from one year to the next. Aaronson, Barrow, and Sander (2007) compared two years of rankings of Chicago public school teachers, based on the teachers' value-added estimates. They found that 57 percent of teachers who were ranked in the top quartile in the first year also ranked in the top quartile in the second year, while 20 percent dropped into the lower half of the quality distribution. Finally, in studying New York City's data reports for teachers with multiple value-added estimates, Corcoran (2010) found that

40 percent of the top 20 percent of teachers in 2007 remained in the top 20 percent in 2008, while 12 percent fell to the bottom 40 percent.

Such uncertainty can be reduced by averaging annual estimates across years. Schochet and Chiang (2010) found that when trying to distinguish a school district's low- or high-performing upper elementary teachers from those with average performance, the rate of misclassification (that is, the rate at which teachers were identified as being better or worse than they actually were)[8] was 36 percent when using only one year of data for each teacher, 26 percent when averaging across three years of data, and 12 percent when using ten years of data.

Appendix B. Creating student achievement variables

To create the four student achievement variables used for this study, the study team followed Washoe County School District rules regarding the inclusion of student achievement scores in teachers' scores, as well as regarding assignment of students to teachers. It was critical to follow these rules because the study examines the stability of teacher scores as they will be derived by the district for use in Nevada's teacher evaluation system. The rules are:

- The scores of students who were not enrolled in their school for the full school year are excluded.
- The scores of students who may have transferred between teachers within a school are included; a student is considered to be assigned to the teacher whose class the student is in at the time of testing.[9]
- A student's scores are included in the computation of a teacher's score if the student's achievement data are coded to the teacher in the dataset, even if the data are coded to more than one teacher.
- The teacher-level scores for teachers who teach more than one class (such as middle school teachers who teach multiple math courses each day) or more than one grade (such as elementary teachers who teach split-grade classes) are derived by pooling all students assigned to the teacher.
- Teacher-level scores (both status scores and growth scores) may be derived only for teachers with at least 10 students who have scores and who have been enrolled in the school for the full school year.[10]

Two additional decision rules follow the Washoe County School District's procedures for teachers who change teaching assignments from one year to the next:

- Teachers who teach different grades within the same school level (elementary or middle school) in different years are included (as teacher scores are assumed to be independent of grade level in the Washoe County School District).[11]
- Teachers who change schools between study years are included, because the student growth percentile model includes no school-level factors.[12]

The number of teachers in the Washoe County School District files for reading and math and the number for whom scores could be computed based on the criteria noted above are presented in table B1. The total number of teachers is the number of teachers in grades 4–8 who had at least one student assigned to them for the subject tested (math or

Table B1. Number of Washoe County School District teachers of reading and math, and number meeting eligibility criteria for the study, by year and across all years

	Math			Reading		
Year	Total number of teachers	Number eligible for study	Percentage eligible	Total number of teachers	Number eligible for study	Percentage eligible
2009/10	664	588	88.6	685	593	86.6
2010/11	677	615	90.8	696	636	91.4
2011/12	662	611	92.3	674	628	93.2
Eligible across the three years	390	369	94.6	404	375	92.8

Source: Authors' analysis of data provided by Washoe County School District.

reading). In each year, at least 86 percent of the teachers were eligible, based on Washoe County School District criteria, to have a teacher-level growth score. In math, 390 teachers met the inclusion criteria for the three years. Of these, 369, or 95 percent, were eligible to have scores computed, and they make up the sample of math teachers in this study. In reading, 404 teachers met the inclusion criteria across all three years of the study. Of these, 375, or 93 percent, were eligible to have scores computed for their class, and they make up the sample of reading teachers in this study.

Appendix C. Design and methods

This appendix provides a brief background about the psychometric theory of generalizability along with details about how it is applied in this study and how it might be applied in similar studies with different assumptions.

Interpretations of a measurement (for example, a test score) taken on a particular day using a particular measurement method are rarely limited to an interpretation of the measurement to that specific day and measurement method. Instead, inferences are made from that single measurement to answer broader questions. For example, a test score can be used to answer whether a student has developed the math knowledge expected of students at his or her grade or whether the student's teacher is a capable teacher of math. These are not questions about performance on a particular day or about performance assessed by a particular method of measurement; they are questions about an individual's enduring traits.

When test scores are interpreted, generalizations move from the particular score observed to a broader universe of possible scores that could have been observed—for example, those that could be achieved on different days, using different measurement methods. Generalizability theory provides a conceptual framework that is useful in accounting for key features, called facets, of the universe of admissible observations. Webb and Shavelson (2005) describe the universe of admissible observations as the collection of observations that would be acceptable to decisionmakers as substitutes for the particular score that was observed. Generalizability theory also provides a statistical method to evaluate the precision of a generalization and to examine how changing the way in which measurements are sampled will alter the precision of inferences.

For this generalizability study of teacher effectiveness measures, the study team used a simple design that includes only one facet, or source of error: the year in which a teacher is measured. The design is shown in table C1. Rows represent teachers, columns represent years, and each cell has one observation, X_{ty}, which is the effectiveness measure observed for teacher t in year y. This is referred to as a crossed design because each teacher is observed each year. In the analysis of the teacher-level growth score, X_{ty} is the median student growth percentile for students of teacher t in year y. In the analysis of the teacher status measure, X_{ty} is the proportion of teacher t's students who were proficient in the subject tested in year y.

Applying a linear model, X_{ty} may be represented as the sum of the expected values of effects associated with rows (teachers), columns (years), their interaction (teachers by years), and random errors. For the single-facet crossed design, the model is expressed for the generalizability study as

$$X_{ty} = \mu + (\mu_t - \mu) + (\mu_y - \mu) + (X_{ty} - \mu_t - \mu_y + \mu) \qquad \text{(C1)}$$

where μ is the grand mean, the expectation of X_{ty} taken over all members of the teacher population and all years in the universe of observations; $(\mu_t - \mu)$ is the effect of teacher t, the deviation from the grand mean of the expected value of the teacher's score taken over all years in the universe of observation; $(\mu_y - \mu)$ is the effect of year y, the deviation from the grand mean of the expected value of the scores for year y taken over all teachers in the population; and $(X_{ty} - \mu_t - \mu_y + \mu)$ is a residual effect, or the portion of the score X_{ty} that is

Table C1. A single-facet crossed design

Teacher	Year					
	1	2	3	4	...	*M*
1	X_{11}	X_{12}	X_{13}	X_{14}	...	X_{1m}
2	X_{21}	X_{22}	X_{23}	X_{24}	...	X_{2m}
3	X_{31}	X_{32}	X_{33}	X_{34}	...	X_{3m}
4	X_{41}	X_{42}	X_{43}	X_{44}	...	X_{4m}
...	...	...	...	...	...	...
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	...	X_{nm}

Source: Authors' construction.

not explained by the other effects. The residual effect includes the effect of the interaction between teachers and years as well as all other random sources of unexplained measurement errors.

The effects may vary. For example, the teacher effect, $(\mu_t - \mu)$, may vary across teachers in the population. Each effect then has an associated variance, which is called a component of variance. The variance component for the effect of teachers is σ_t^2, the variance component for the effect of years is σ_y^2, and the variance component for the residual effect is $\sigma_{ty,e}^2$. The variance of X_{ty}, taken over all teachers in the population and all years in the universe of observations, is the sum of the three variance components. Just as the variance of X_{ty} is the variance of scores taken in a single year, variance components are also associated with one year. This is important when the variance components are used to estimate the effects of changing the number of years of data that enter a teacher's score.

Variance components are the parameters estimated in generalizability theory analyses. (For details about the methods to estimate variance components, see Brennan [2001] or Cronbach, Gleser, Nanda, and Rajaratnam [1972].) With variance components, it is possible to identify how variance in observed scores is expected to be affected by different facets of the universe and by sampling different numbers of observations from each facet. It also becomes possible to examine how the variance in scores would be affected by different types of designs for data collection. As a result, it is possible to consider how to maximize the information from a score and minimize the impact of other sources of variance and, thus, to design data collection plans that provide dependable measurements.

A key concept in generalizability theory is that, unlike in classical test theory, there is not a single "true" score or "error" score. In classical test theory a person's observed score is simply the sum of a true score and an error score. Generalizability theory recognizes multiple sources of influence on scores. Which of those influences enter the true score and which enter the error score depend on the use of the score and how it is to be interpreted.

In the case of the simple single-facet crossed design in this study, there are three effects, each having an associated variance component: residual, teacher, and year. The residual component is equivalent to the error variance of classical test theory. The teacher component is equivalent to the true score variance of classical test theory. The year component does not have a comparable term in classical test theory because classical test theory assumes it to be zero (that is, it assumes strictly parallel forms of tests or measurement methods, with each form having the same mean value of scores), while generalizability

theory relaxes that assumption. For the teacher-level status measure (derived from student scores on the statewide achievement test), the variance component for year would be greater than zero if, for example, there had been a change in the statewide test that resulted in the test becoming easier, on average, than previous years' tests. That type of change may be a source of error for some decisions but not for others. It would be a source of error when the scores of teachers are used to make a decision that involves an absolute comparison, such as the comparison of a teacher's score to a cutscore that teachers must meet in order to be classified as effective. A change in the average level of difficulty of the test from one year to the next could change a teacher's position relative to the criterion, even in cases where the teacher's effectiveness had not changed. In contrast, the change in test difficulty would not be a source of error for decisions involving relative comparisons, such as a decision to select the top 10 percent of the teachers for awards. The rank order of teachers would not be affected by a shift in test difficulty that adds a constant to each teacher's score.

This study assumes an absolute comparison, and the standard error of measurement includes both the year and residual variance components:

$$SEM_{abs} = \sqrt{\frac{\sigma^2_{y}}{n_{y}} + \frac{\sigma^2_{ty,e}}{n_{y}}} \qquad \text{(C2)}$$

where n_y is the number of years of data that are averaged to obtain the teacher's scores. The standard error of measurement is the square root of the total error variance, and it provides a measure of the error in units of the scale of the teacher score. Since it is reported in units of the score scale, the magnitude of the standard error of measurement cannot be judged independent of that scale. The standard error of measurement is useful in constructing confidence bands or intervals that, with a particular level of certainty, are likely to include the true (error-free) score of a teacher. If using a score from a single year of data, the standard error of measurement is simply based on the sum of the two variance components. The standard error of measurement is reduced (and reliability increased) if two or more years of data are sampled and averaged to obtain a teacher's score.

Some states may have systems using a relative comparison, and the standard error of measurement in those cases includes only the residual component:

$$SEM_{rel} = \sqrt{\frac{\sigma^2_{ty,e}}{n_{y}}} \, . \qquad \text{(C3)}$$

The generalizability coefficient ranges from 0 to 1 and is analogous to the reliability coefficient of classical test theory in that it is defined as a ratio of the variance among teachers to the total variance. For absolute decisions, as used in this study, the generalizability coefficient is[13]

$$\rho^2_{abs} = \frac{\sigma^2_{t}}{\sigma^2_{t} + \frac{\sigma^2_{y}}{n_{y}} + \frac{\sigma^2_{ty,e}}{n_{y}}} \, . \qquad \text{(C4)}$$

For relative decisions, the generalizability coefficient is

$$\rho^2_{rel} = \frac{\sigma^2_t}{\sigma^2_t + \frac{\sigma^2_{ty,e}}{n_y}}. \qquad \text{(C5)}$$

In equations C2–C5 the term n_y is the number of years of data that enter the teacher's score. By sampling more years, one would expect to reduce the standard error of measurements and increase the reliability coefficients. Thus, once the components of variance have been estimated, the standard error of measurements and generalizability coefficients can be estimated for situations that differ in the number of years of data that are averaged to obtain a teacher's score. Thus, equation C4 is the basis for the estimates reported in this study for different numbers of years of data.

The GENOVA software package developed by Brennan (2001) was used to estimate the three components of variance (teachers, years, and the variance unexplained by these two sources) and to compute two indicators of stability: the generalizability coefficient and the standard error of measurement.

Appendix D. Calculating misclassifications of effectiveness

The misclassification example in the findings section identifies the proportion of misclassifications for a "typical" math teacher whose true score was 50 when the cutscore for being considered effective was 40. If the true score for each teacher were known, the misclassification rate for each teacher could be calculated using the same method, and the rates could be averaged across teachers to obtain the expected misclassification rate for the group as a whole. For teachers with a true score above 40, the likelihood of misclassification would be the proportion of scores that fall below 40, as in the previous example. For teachers with a true score below 40, misclassification would occur when errors caused scores to fall above 40.

While a teacher's true score cannot be measured, it can be estimated. Using the three years of scores provided for each teacher, along with the study findings, the study team estimated teachers' true scores for every math teacher in the sample and then computed the proportion of classifications of teachers that would be in error when teachers were classified as ineffective or effective against a cutscore of 40.[14] The study team looked at two types of classification errors: identifying teachers with a true score above the cutscore as ineffective and failing to identify ineffective teachers with a true score below the cutscore.

When classifications are based on one year's annual teacher-level growth score, the proportion of teachers expected to be misclassified is .14 (table D1).[15] The proportion of effective teachers whose true growth score is 40 or higher and who are likely to be incorrectly classified as ineffective is .13. The proportion of ineffective teachers whose true growth score is below 40 and who are likely to be incorrectly classified as effective is .42. While the latter proportion is high, it represents fewer teachers than the proportion of misclassified effective teachers. Some 17 teachers had a true score below 40, so an error rate of .42 means that 7 teachers would likely be misclassified. By comparison, 352 teachers had a true score of 40 or higher, so a misclassification rate of .13 means that 46 teachers would likely be misclassified.

The expected misclassification rates decline when teacher-level growth scores are derived by averaging two or more annual scores (see table D1).

The methods used in this example can also be used to examine how changes in the cutscore will alter expected misclassification rates.

Table D1. Expected misclassification rates when identifying ineffective teachers as those with a growth score estimate below 40

Number of years of growth scores included in the estimate	All teachers (n = 369)	Effective teachers (n = 352)	Ineffective teachers (n = 17)
1	.14	.13	.42
2	.10	.09	.39
3	.08	.06	.37

Source: Authors' analysis of 2009/10, 2010/11, and 2011/12 data provided by Washoe County School District.

Notes

1. Collins and Amrein-Beardsley (2014) surveyed all 50 states and the District of Columbia to learn whether they were using or piloting a student growth percentile model or a value-added model and, if so, which one. The District of Columbia and 22 of the 40 states that reported using or developing a model identified their model. Of these, 13 were using or piloting the student growth percentile model. The other 18 of the 40 states indicated that they were using or developing a growth model or a value-added model but did not identify the model. New Jersey Department of Education (2012) indicated that as many as 20 states may be considering the student growth percentile model.
2. The stability of a measure over time (the focus of this study) provides information about the measure's reliability (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Haertel, 2006). Within an argument-based approach to validity (Kane, 2006, 2013), this information is also recognized as providing evidence to evaluate the validity of the inferences made in the measure's interpretations.
3. In creating the comparison group for a student, the analysis takes into account as many prior years of achievement test scores as the student has. Grade 4 students, for example, would have, at most, one prior year of scores because in Nevada, achievement testing begins in grade 3. Grade 8 students would have, at most, five prior years of scores. The grouping of students by their achievement histories is not exact—that is, rather than looking for students with exact matches in prior scores, the statistical method known as regression analysis is used to form approximate groups with similar, but not exact, achievement histories. More information about the method can be found in Castellano and Ho (2013), as well as in the original citation, Betebenner (2011).
4. As of 2014, at least 50 percent of a Nevada educator's evaluation was to be based on student achievement data from the state's accountability system, with 45 percent of the total evaluation based on scores derived from the student growth percentile model. New legislation passed in spring 2015 reduced to 40 percent the weight of student data in a teacher's evaluation and further specified that half of that 40 percent would come from district rather than state tests, without specifying the proportion to be based on scores from the student growth percentile model. Because of changes to the statewide testing program, no student achievement data will be included in teacher evaluations for 2015/16, and for 2016/17, 20 percent of the evaluations will be based on student achievement data, with half coming from state and half from district tests. The new 40 percent requirement will be in full effect starting in 2017/18.
5. The coefficient and standard error of measurement can be estimated for any number of years of data by applying the statistics derived from the data and using standard assumptions from psychometric theory. Estimates are presented for one, two, and three years of data so that educators and policymakers can see how the values change as years of data are added. More than three years seems longer than policymakers or administrators would want to wait to make a decision about a teacher's effectiveness. Teachers in Nevada achieve tenure in three years; thus, a major decision that might be based on the scores examined in this report can use, at most, three years of data.
6. Cutscores are selected points on the score scale of a test that are used to determine whether a particular test score is sufficient for some purpose; for example, student performance on a test may be classified into one of several categories such as basic,

proficient, or advanced based on cutscores (see, for example, Zieky & Perie, 2006). At the time of this study, Nevada has not yet set cutscores for its teacher evaluation system. The study team selected a score of 40 as an example because it had been discussed by the Washoe County School District as a possible cutscore. This example shows how classification errors might be examined for a particular cutscore and, in doing so, demonstrates how the standard error of measurement could be used to evaluate different error rates for different cutscores during the design of an evaluation system.

7. At the time of the study by Lash et al. (2013), Nevada had not determined a cutscore, or criterion score, to identify low-performing schools for the state's school accountability or principal evaluation system. Policymakers were discussing setting the cutscore at 40.
8. Schochet and Chiang (2010) explain that classification error, in their context, relates to the false positive (Type I) and negative (Type II) error rates from classical hypothesis testing. The Type I error rate is essentially the probability that the test of teacher effectiveness will erroneously find that a truly average teacher performed significantly worse than average—that is, the probability that an average teacher will be erroneously identified for remediation. Conversely, the false negative error rate is the probability that the test will fail to identify teachers whose true performance is a certain number of standard deviations below average—that is, the probability that a low-performing teacher will not be identified for remediation even though he or she warrants it.
9. The district follows this policy because its experience is that the data about transfers are unreliable and that there are few within-school transfers after the first few weeks of school.
10. Computation of the student growth score, a student growth percentile, requires that a student have a test score in the current year and at least one test score in a previous year. Students who have a student growth percentile, then, will have a proficiency score for the current year. Students with proficiency scores may be missing a student growth percentile if they were never tested previously in Nevada. However, even students new to the Washoe County School District will have a student growth percentile if they attended school in Nevada in the past, because the Nevada Department of Education computes student growth percentile scores in an analysis that pools students from all districts in the state.
11. The study excluded a few teachers who changed school levels because the district was interested in analyzing the data separately by school level in the future and wanted the samples for those future analyses to contain the teachers in the current sample. For this reason, three teachers were excluded from the analysis of math scores and four from the analysis of reading scores.
12. Omitting school effects from a model estimating teacher effects attributes any school effect (contextual, direct, or indirect) to teachers. Such effects tend to be left out of value-added models for several reasons. For example, it is hard to determine how a teacher's principal or colleagues may have influenced a teacher's score (Corcoran, 2010) or how schools are selected by students and teachers (Hanushek & Rivkin, 2010), and research suggests that the between-school variance in teacher value-added models tends to be fairly small compared to the variance within schools (Hanushek & Rivkin, 2010).
13. For absolute decisions, Brennan (2001) refers to the coefficient as an index of dependability rather than as a generalizability coefficient. For the sake of simplicity, this report refers to the coefficients derived for both absolute and relative decisions as a reliability coefficient.

14. A regression equation known as Kelley's formula provides a means to estimate the true score for each teacher (Hubert & Wainer, 2013): $\hat{T}_t = \overline{X} + r_{xx'}(X_t - \overline{X})$, where $\hat{T}_t$ is the estimated true score for teacher *t*, $\overline{X}$ is the mean of teacher scores, $r_{xx'}$ is the coefficient summarizing the stability of the score, and X_t is the observed score for teacher *t*. For this example the average of the three estimates of a teacher's growth (rather than one of them) was used as the observed score, and thus the coefficient is for the average of three annual scores.
15. As noted, the misclassification results assume a normal distribution of errors with a standard deviation equal to the standard error of measurement. They represent the proportion of observed scores that would fall below the relevant cutscore for teachers whose true scores actually lie above the cutscore, and vice versa—that is, the proportion that would be expected to be judged incorrectly due to measurement error.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labor Economics, 25*(1), 95–135.

American Educational Research Association. (November, 2015). *AERA issues statement on the use of value-added models in evaluation of educators and educator preparation programs*. Washington, DC: Author. Retrieved December 21, 2015, http://www.aera.net/Newsroom/NewsReleasesandStatements/AERAIssuesStatementontheUseofValue-AddedModelsinEvaluationofEducatorsandEducatorPreparationPrograms/tabid/16120/Default.aspx.

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.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author. Retrieved June 19, 2014, from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf.

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*(1), 1–31. http://eric.ed.gov/?id=EJ1020233

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. Dover, NH: National Center for the Improvement of Educational Assessment.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Castellano, K. E., & Ho, A. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers. http://eric.ed.gov/?id=ED551292

College Board. (2013). *Test characteristics of the SAT: Reliability, difficulty levels, completion rates*. New York, NY: Author. Retrieved February 20, 2014, from http://media.collegeboard.com/digitalServices/pdf/research/Test-Characteristics-of-SAT-2013.pdf.

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1), 1–32. http://eric.ed.gov/?id=EJ1020222

Corcoran, S. (2010). *Can teachers be evaluated by their students' test scores? Should they be?* (Report for the Education Policy for Action Series). Providence, RI: Annenberg Institute for School Reform at Brown University. http://eric.ed.gov/?id=ED522164

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley & Sons.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2011). *Getting teacher evaluation right: A brief for policymakers.* Capitol Hill Research Briefing convened by the American Educational Research Association and the National Academy of Education. Washington, DC. http://eric.ed.gov/?id=ED533702

Davidson, A. H., Ozdemir, S., & Harris, J. (2010, May). *An approach to criterion-referenced growth modeling: One state's application of student growth percentiles.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems.* Washington, DC: Brown Center on Education Policy at the Brookings Institution. http://eric.ed.gov/?id=ED518919

Goldschmidt, P., Choi, K., & Beaudoin, J. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Washington, DC: Council of Chief State School Officers. http://eric.ed.gov/?id=ED542761

Good, T. L. (2014). What do we know about how teachers influence student performance on standardized tests and why do we know so little about other student outcomes? *Teachers College Record, 116*(1), 1–31. http://eric.ed.gov/?id=EJ1020231

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2014). *A comparison of growth percentile and value-added models of teacher performance (Working paper # 39).* East Lansing, MI: The Education Policy Center at Michigan State University. http://eric.ed.gov/?id=ED558130

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (pp. 65–110). Westport, CT: Praeger Publishers.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores* (14th William H. Angoff Memorial Lecture). Princeton, NJ: ETS.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers and Proceedings, 100*(2), 267–271.

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice*, *10*(2), 33–41. Retrieved August 1, 2014, from http://ncme.org/linkservid/6606715E-1320–5CAE-6E9DDC581EE47F88/showMeta/0/.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1), 1–28. http://eric.ed.gov/?id=EJ1020230

Hubert, L., & Wainer, H. (2013). *A statistical guide for the ethically perplexed.* Boca Raton, FL: Taylor & Francis Group/CRC Press.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. http://eric.ed.gov/?id=EJ996447

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record, 116*(1), 52. http://eric.ed.gov/?id=EJ1020224

Lash, A., Peterson, M., Vineyard, R., Barrat, V., & Tran, L. (2013, April). *The generalizability of school growth scores derived from student growth percentiles for use in school accountability and principal evaluation systems.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606. http://eric.ed.gov/?id=EJ863346

National Council of State Boards of Nursing. (n.d.). *Quarterly examination statistics: Reliability of the NCLEX examinations*. Retrieved May 12, 2014, from http://www.ncsbn.org.

Nevada Department of Education. (2010). *Nevada Growth Model of Achievement (NGMA) rollout*. Carson City, NV: Author. Retrieved December 2, 2010, from http://doe.nv.gov/Accountability/NevadaGrowthModel_of_Achievement-Rollout.pdf.

New Jersey Department of Education. (2012). *An illustration of SGP adoption in the United States*. Trenton, NJ: Author. Retrieved May 12, 2012, from http://www.state.nj.us/education/njsmart/performance/SGP_Adoption.pdf.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 1–22. http://eric.ed.gov/?id=EJ913473

Salkind, N. J. (2008). *Encyclopedia of educational psychology*. Thousand Oaks, CA: SAGE Publications, Inc.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE No. 2010–4004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://eric.ed.gov/?id=ED511026

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE Publications.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (pp. 43–66). Hoboken, NJ: John Wiley & Sons.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 717–719). Hoboken, NJ: John Wiley & Sons.

Zieky, M. J., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.

The Regional Educational Laboratory Program produces 7 types of reports

Making Connections
Studies of correlational relationships

Making an Impact
Studies of cause and effect

What's Happening
Descriptions of policies, programs, implementation status, or data trends

What's Known
Summaries of previous research

Stated Briefly
Summaries of research findings for specific audiences

Applied Research Methods
Research methods for educational settings

Tools
Help for planning, gathering, analyzing, or reporting data or research