American Educational Research Journal
December 2019, Vol. 56, No. 6, pp. 2116–2146
DOI: 10.3102/0002831219841369

The Consistency of Composite Ratings of Teacher Effectiveness: Evidence From New Mexico

Sy Doan
Vanderbilt University
Jonathan D. Schweig
Kata Mihaly
RAND Corporation

Contemporary teacher evaluation systems use multiple measures of performance to construct ratings of teacher quality. While the properties of constituent measures have been studied, little is known about whether composite ratings themselves are sufficiently reliable to support high-stakes decision making. We address this gap by estimating the consistency of composite ratings of teacher quality from New Mexico's teacher evaluation system from 2015 to 2016. We estimate that roughly 40% of teachers would receive a different composite rating if reevaluated in the same year; 97% of teachers would receive ratings within ±1 level of their original rating. We discuss mechanisms by which policymakers can improve rating consistency, and the implications of those changes to other properties of teacher evaluation systems.

Keywords: teacher accountability, performance assessment, evaluation, simulation, measurement

---

Sy Doan is a PhD candidate at Vanderbilt University's Peabody College of Education and Human Development, 230 Appleton Place, Nashville, TN 37203; e-mail: *sy.doan@vanderbilt.edu*. His research focuses on teacher evaluation policy and the measurement of teacher effectiveness.

Jonathan D. Schweig, PhD, is a social scientist at the RAND Corporation. His research focuses on education policy and the measurement of instructional practices, classroom and school climate, and social and emotional competencies.

Kata Mihaly, PhD, is a senior economist and associate director for the Economics, Sociology, and Statistics Department at the RAND Corporation. Her research focuses on the use of multiple measures to evaluate teachers, and the evaluation of programs designed to help teachers improve their practice.

Multiple measure systems are the predominant form of teacher evaluation in the post-No Child Left Behind era (Steinberg & Donaldson, 2016). By combining data from classroom observations, student test scores, student surveys, and other measures of teaching quality and effectiveness, these systems produce ratings of teacher quality that can be more robustly predictive of student outcomes than when these measures are used in isolation (T. J. Kane, McCaffrey, Miller, & Staiger, 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Accordingly, states and districts increasingly use composite ratings to inform the compensation, professional development, and dismissal of teachers. Adopting high-stakes multiple measure systems may positively affect subsequent teacher performance and student test scores (Cullen, Koedel, & Parsons, 2016; Dee & Wyckoff, 2015).

Effective teaching is complex and multifaceted (Borko, 2004; Cochran-Smith, 2003; Leinhardt & Greeno, 1986), making the use of multiple measures to evaluate teachers intuitively appealing. However, best practices for using multiple measure systems to inform high-stakes decisions remains elusive. A broad literature exists on the validity and reliability of the most popular components of multiple measure systems, for example, value-added estimates (Koedel, Mihaly, & Rockoff, 2015), classroom observations (Ho & Kane, 2013), and student surveys (English, Burniske, Meibaum, & Lachlan-Haché, 2016). Fewer studies examine how these components operate in concert when used as parts of a composite score and whether the resulting composites are sufficiently precise to support the human capital decisions being made using them (Martínez, Schweig, & Goldschmidt, 2016; Steinberg & Kraft, 2017).

This article addresses this gap in the literature by being among the first to estimate the consistency of summative ratings of teacher effectiveness using data obtained from an active, at-scale teacher evaluation system. Specifically, we adapt simulation methods previously used in studies of student assessments to estimate the consistency of teacher effectiveness decisions based on the multiple measures employed in the New Mexico Educator Effectiveness System (NMTEACH) during the 2015–2016 academic year, defining consistency as the likelihood that a teacher would receive the same NMTEACH rating if the evaluation process were repeated in the same school year. Our analysis addresses the following research questions:

Research Question 1: To what extent are NMTEACH ratings of teacher quality consistent across simulated repetitions of the evaluation process?

Research Question 2: To what extent can rating consistency potentially be improved by changes in evaluation policy? Specifically, to what extent does rating consistency vary as a function of (a) the reliabilities of the component measures, (b) weights assigned to the component measures, and (c) the locations of the cut-points between rating categories?

Results for Research Question 1 provide estimates of the consistency of NMTEACH ratings under "business-as-usual" conditions while the analyses conducted under Research Question 2 demonstrate how changes to component measure and cut-point properties can affect composite rating consistency, suggesting possible policy levers for changing these properties. This study makes two contributions to the literature. First, estimates from this study provide important empirical benchmarks for the level of consistency policymakers can expect when implementing and revising teacher evaluation policy. Second, the empirical approach can be adapted and used to explore classification consistency in districts and states across the country.

We first provide a synopsis of existing literature on multiple measure teacher evaluation and a description of NMTEACH. Next, we describe our methods for estimating the consistency of teacher ratings, using both parametric (Douglas & Mislevy, 2010; Martínez et al., 2016) and nonparametric (Brennan & Wan, 2004) techniques. Last, we present results and discuss their substantive implications regarding the relationship between policy design and rating consistency.

Background

The Case for Multiple Measure Teacher Evaluation

The current proliferation of multiple measure teacher evaluation was preceded by a swell of advocacy and research arguing that teacher evaluation systems failed to provide useful feedback on teacher performance (Bill & Melinda Gates Foundation, 2010; Weisberg et al., 2009). Traditionally, teacher evaluation consisted of loosely structured administrator observations, which critics claimed failed to discriminate among teachers of different effectiveness levels or support teacher professional development (Weisberg et al., 2009; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1985). Concurrently, research suggested that teachers were the largest within-school contributor to student achievement (Goldhader, Brewer, & Anderson, 1999; Nye, Konstantopoulos, & Hedges, 2004) and that teacher background characteristics or qualifications inadequately explain their contributions (Baker et al., 2010; Nye et al., 2004; Rivkin, Hanushek, & Kain, 2005).

In response to these concerns, several Obama Administration-era policies, such as Race to the Top, No Child Left Behind waivers, and Teacher Incentive Fund grants, either encouraged or required states and districts to develop multiple measure teacher evaluation systems that combined information from administrator observations with outcome-based measures of effectiveness, such as student test scores or survey responses (U.S. Department of Education, 2010). Relative to single measure systems, multiple measure systems are thought to provide more valid (Baker et al., 2010; Goe, Holdheide, & Miller, 2011) and stable (Steele, Hamilton, & Stecher,

2010) ratings of teacher performance. From 2010–2011 to 2016–2017, 46 states enacted reforms of their teacher evaluation systems, with roughly 80% (36 of 46) of these systems requiring that measures of student test score performance be incorporated alongside classroom observation measures (Kraft & Gilmour, 2017; Steinberg & Donaldson, 2016).

Research on Individual Teacher Evaluation Measures

The base of knowledge on the properties of the constituent measures of multiple measure systems (e.g., value-added models, classroom observations, student surveys) is rapidly growing. There have been several high-profile experimental and quasiexperimental studies showing that value-added estimates of teacher effectiveness are minimally biased predictors of student test score achievement (e.g., T. J. Kane et al., 2013) and are associated with long-run student outcomes (Chetty, Friedman, & Rockoff, 2014). However, evidence pointing to value-added measures' relatively low reliability, incentives for "teaching to the test," and lack of clear guidance for teachers' formative development raise concerns regarding the appropriateness of their use for evaluating individual teachers (American Educational Research Association, 2015; Corcoran & Goldhaber, 2013).[1]

There is also evidence that evaluation based on structured teacher observations, such as the Charlotte Danielson Framework for Teaching; Danielson, 2007) and CLASS (Pianta, Hamre, Haynes, Mintz, & La Paro, 2006), can be used to improve teacher practice (Taylor & Tyler, 2012). Observation scores are also positively associated with student achievement (e.g., Bacher-Hicks, Chin, Kane, & Staiger, 2017; T. J. Kane, Taylor, Tyler, & Wooten, 2010). However, observations also have many limitations. At scale, they are expensive and labor intensive (Rothstein & Mathis, 2013) and obtaining reliable scores from observation rubrics may pose significant administrative challenges (Hill, Umland, Litke, & Kapitula, 2012; Ho & Kane, 2013). Additionally, some research shows that observation scores are potentially biased by observer (Grissom & Loeb, 2017; Ho & Kane, 2013) and student (Mihaly & McCaffrey, 2014; Steinberg & Garrett, 2016) characteristics.

Finally, recent work has shown that indicators derived from student surveys, such as the Panorama Student Survey (Panorama, 2015) and the Tripod Student Survey (Ferguson, 2010), also correlate significantly with student achievement (Kyriakides, 2005), value-added and observation-based ratings of instructional practice (Bill & Melinda Gates Foundation, 2010), and can be used to distinguish reliably between the practices of different teachers (Balch, 2012; Ferguson 2010). However, student ratings can be susceptible to halo effects (Wallace, Kelcey, & Ruzeck, 2016) and influenced by student demographics, age, and similar factors (Cherng & Halpin, 2016; Ferguson, 2010; Worrell & Kuterbach, 2001).

Research on Multiple Measure Systems

Research on the properties of composite, rather than individual, evaluation measures, is relatively lacking. Several papers examine correlations between observation scores and value-added estimates, finding consistently significant relationships of varying strength (Grossman, Loeb, Cohen, & Wyckoff, 2013; Strunk, Weinstein, & Makkonen, 2014). While these studies establish concurrent validity between measures, they do not provide information on the properties of composite scores that combine these measures. It is often assumed that combining measures produces more precise estimates of teacher quality, thus improving decision making (Jackson & Mackler, 2016). However, there is abundant methodological literature challenging this conventional wisdom (e.g., Cronbach, Linn, Brennan, & Haertel, 1997). Aggregating multiple measures does not, in general, "cancel out" the measurement error inherent in each individual measure. Rather, the reliability of composites depends on the component weights, reliabilities, variances, and correlations of the individual measures (M. Kane & Case, 2004). Composite scores can be less precise than individual measures when more reliable measures are combined with less reliable ones (M. Kane & Case, 2004). Given that combination rules have been demonstrated to affect the validity and reliability of composites scores (Chester, 2003; Douglas & Mislevy, 2010; Martínez et al., 2016), research explicitly studying composites, rather than simply the relationships between their constituent measures, is needed.

Research using the data from the Measures of Effective Teaching (MET) project are among the most thorough investigations of the properties of individual and composite measures of teacher quality. This research demonstrates that observation, value-added, and student survey information can be combined to produce unbiased composite measures of teacher effectiveness that are predictive of multiple student outcomes (T. J. Kane et al., 2013; Mihaly et al., 2013). Two recent papers describe how changes to combination rules affect teacher ratings. Martínez et al. (2016) consider how the use of different combination models (e.g., compensatory, disjunctive, conjunctive) affect teachers' "pass" rate and the consistency of those ratings. Steinberg and Kraft (2017) focus their attention on compensatory models, which weight and combine individual teacher evaluation measures to form composite scores of teacher effectiveness, finding that changes to measure weights can substantially affect the final distribution of composite effectiveness ratings.

The Importance of Consistency in Teacher Evaluation

Much of the prior research on teacher evaluation measures focuses on properties of the summative scores themselves, including their predictive validity, or the extent to which differences in these measures are predictive

of differences in the outcomes of the students taught by those teachers. By comparison, far less focuses on whether these measures can be used to consistently classify teaching as effective or ineffective, a critical property for the fairness and efficiency of a teacher evaluation system.

While teacher evaluation measures are intended to capture different dimensions of teaching effectiveness, scores obtained from these measures are, to varying degrees, affected by several factors unrelated to effectiveness. For example, if determinations about teaching effectiveness are made based on classroom observations, an individual teacher's scores may depend on the specific days when observations occur, the specific students in the classroom, the specific lesson being taught, or the severity or leniency of the observer, among other factors (Cronbach et al., 1997; M. Kane, 2011). This measurement error introduces uncertainty into summative scores and, as a result, into decisions about whether to classify teaching as effective or ineffective. If decisions about effectiveness are heavily influenced by factors outside of a teacher's control, stakeholder perceptions of evaluation system fairness can be weakened, hampering the efficiency with which policymakers can use rewards and sanctions based on these scores to develop a more effective teaching workforce. However, if decisions about effectiveness are relatively robust to this uncertainty, there is more evidence to support teacher evaluation claims.

Evidence of the extent to which classification decisions are robust to measurement error is often referred to as classification consistency. In this analysis, we estimate the within-year classification consistency of teachers' NMTEACH ratings and investigate how different features of the NMTEACH system affect consistency. In doing so, we provide policymakers and researchers with accessible benchmarks with which to gauge the reliability of the *decisions* based on summative scores derived from multiple measure systems commonly used to evaluate teachers across the United States.

Our analysis builds on prior work studying multiple measure evaluation systems in two ways. We are the first to study the properties of ordinal effectiveness ratings using data collected from an active, at-scale teacher evaluation system, presenting findings that are arguably more generalizable than prior studies using data from research-based settings such as the MET project. Second, our simulation-based methods allow us to look expressly at how measurement error in the underlying NMTEACH measures affect the consistency of the resulting composite ratings, offering policymakers and program designers insight into the potential gains (or losses) to rating consistency that could result from changes to component measure reliability.

The NMTEACH System

The NMTEACH Educator Effectiveness System ("NMTEACH") was implemented during the 2013–2014 academic year as the state of New

Mexico's teacher evaluation policy for all public and state-run charter schools. NMTEACH uses a compensatory model of teacher evaluation broadly similar to other compensatory systems around the country: Teachers are assigned an annual Level 1–5 rating on the basis of a composite summative score constructed from their observation scores, value-added estimates (if available), and other measures such as teacher attendance and parent/student survey results. Teachers classified as Level 1 ("Ineffective") or Level 2 ("Minimally Effective") are assigned to a "professional growth plan," with districts retaining ultimate discretion for assigning professional growth plans. Additional detail on the measures and composite score calculation is provided below.

Measures

NMTEACH incorporates an expansive set of teacher evaluation measures, relative to other multiple measure systems across the country (Kraft & Gilmour, 2017). During the 2015–2016 academic year, NMTEACH used teacher performance on up to five measures, or components, to calculate teachers' composite scores and ratings (1) teachers' "overall" value-added score or VAS, (2) observation scores from Domains 2 and 3 of the NMTEACH observation rubric, (3) observation scores from Domains 1 and 4 of the NMTEACH observation rubric, (4) teacher attendance, and (5) parent/student surveys.

Teacher VASs are estimated separately by grade, subject, and year using a teacher random effects model controlling for up to 2 years of prior test scores, whether a student was in an intervention course (e.g., ESL [English as second language], reading intervention), and the proportion of the academic year a student was enrolled in that teacher's course. To create a summative NMTEACH score, an "overall" VAS is calculated for each teacher, which is a student-count weighted average of all available year-grade-subject specific VAS over the past 3 years. The NMTEACH classroom observation protocol is a modified version of the Charlotte Danielson Framework for Teaching rubric and consists of four domains: (1) Preparation and Planning, (2) Creating an Environment for Learning, (3) Teaching for Learning, and (4) Professionalism. Teachers' scores in Domains 2 and 3 and Domains 1 and 4 are weighted separately under NMTEACH. Starting in 2015–2016, teachers clearing a specific benchmark on their overall and VA scores are only required to be evaluated once annually on Domains 2 and 3; the remaining teachers are evaluated 2 or 3 times a year, depending on the specifics of their district plan. Nearly all teachers were evaluated on Domains 1 and 4 only once. Student (for Grade 3–12 teachers) and parent (for Grade K–2 teachers) surveys are factored into teachers' summative scores. These surveys contain 10 items asking students/parents to rate their teacher's ability to create "opportunities to learn." The use of the surveys and teacher attendance was left to the discretion of the district.

Table 1
Measure Weight Allocations, by Step

Step	VAS, %	Domains 2 and 3, %	Domains 1 and 4, %	Attendance, %	Surveys, %
1 (0 years test data)	0	50	40	5	5
2 (1–2 years test data)	25	40	25	5	5
3 (3+ years test data)	50	25	15	5	5

Note. VAS = value-added score. Table 1 presents measure weights (as percentage of summative score) for the three steps used in NMTEACH (New Mexico Educator Effectiveness System) in 2015–2016. Step assignment is based on the years of valid student test data a teacher has, that is, Step 1 teachers have 0 years, Step 2 teachers have 1–2 years, and Step 3 teachers have 3+ years of test data.

Composite Scores

NMTEACH produces a 22–200 composite score by taking a weighted sum of teachers' performance on each of the component measures. Teachers' 22–200 composite scores are then used to assign a 1–5 NMTEACH summative rating (or effectiveness level). To place all measures on a common scale, raw scores in each component are transformed into a proportion (or proportion-like) 0–1 score. For observations and student surveys, a proportion is calculated by dividing a teacher's points earned by total possible points, summed across all nonmissing indicators/items. Teachers' attendance rates, already proportions, require no further transformation. VAS cannot be transformed into a proportion of total possible points since they are bounded by negative and positive infinity. Therefore, teachers' overall VAS are converted to percentiles using a cumulative distribution function to obtain a proportion-like measure. This conversion results in teachers with median VAS earning 50% of total points possible.

Once scaled values are obtained for each component, they are multiplied by the total points possible for that component and summed to determine a teacher's final 22–200 composite score. The set of weights (referred to as "Steps" within the NMTEACH framework) assigned to each teacher's component measures vary according to the number of years for which a teacher has valid student achievement data. Table 1 describes the weights assigned to each measure, disaggregated by step.

Table 2 shows the ratings and their associated summative score cutpoints. The use of five rating categories is relatively rare compared to other states that use multiple measure teacher evaluation systems, where four categories are more common (Kraft & Gilmour, 2017).

Table 2
2015–2016 NMTEACH Ratings

Rating	Lower Bound	Upper Bound
5. Exemplary	173	200
4. Highly effective	146	<173
3. Effective	119	<146
2. Minimally effective	92	<119
1. Ineffective	22	<92

Note. NMTEACH = New Mexico Educator Effectiveness System.

Data and Methods

Data

Our analysis uses data from the 2015–2016 academic year. During that year, roughly 21,000 New Mexico teachers received an NMTEACH effectiveness level. Of these 21,000 teachers, 17% of teachers were evaluated using Step 1 weights, with the remaining 83% evenly split between Step 2 and Step 3 teachers. Table 3 presents descriptives for NMTEACH measure scores and demographic characteristics, disaggregated by step. Measure scores are presented in "proportion" form so that they are comparable across teachers in different steps; histograms for individual measure scores are available in Appendix Figure A1, in the online version of the journal. Across all three steps, teachers earned roughly 70% of available points on all measures except for VAS, where the average Steps 2 and 3 teacher, by construction, earned roughly 50% of total possible points.

Table 4 shows the proportion of teachers using each measure toward their summative score, by step. While the proportion of teachers using value-added estimates and Domains 2 and 3 scores are as expected, there is variation in the use of Domains 1 and 4 scores, teacher attendance, and survey measures, with over 92% of teachers receiving Domains 1 and 4 scores, and fewer districts choosing to use teacher and student surveys.

In Table 5, we present the distribution of NMTEACH ratings and average points earned during the 2015–2016 school year, disaggregated by step. As has been noted elsewhere, NMTEACH is one of few teacher evaluation systems that produce roughly normally distributed summative ratings (Kraft & Gilmour, 2017). This is broadly true of the ratings in all steps, though the concentration of Level 3 ratings is highest among Step 1 teachers and becomes more diffuse as steps increase in weight allotted to VAS. Additionally, Step 1 teachers, on average, score 10 and 15 points higher than Step 2 and Step 3 teachers, respectively.

Table 3
Descriptive Characteristics

	Step		
Measures	1	2	3
VAS	—	0.504 (0.257)	0.529 (0.236)
Domains 2 and 3	0.702 (0.099)	0.698 (0.1)	0.73 (0.101)
Domains 1 and 4	0.712 (0.114)	0.711 (0.111)	0.745 (0.115)
Parent and student surveys	0.799 (0.105)	0.803 (0.111)	0.817 (0.103)
Teacher attendance	0.755 (0.246)	0.782 (0.228)	0.776 (0.234)
Black	0.015 (0.123)	0.016 (0.125)	0.011 (0.103)
Hispanic	0.315 (0.465)	0.33 (0.47)	0.349 (0.477)
Other	0.065 (0.247)	0.046 (0.209)	0.044 (0.205)
White	0.604 (0.489)	0.608 (0.488)	0.597 (0.491)
Associates	0.003 (0.056)	0.001 (0.034)	0 (0.018)
Bachelors	0.566 (0.496)	0.586 (0.493)	0.548 (0.498)
Doctorate	0.008 (0.092)	0.008 (0.087)	0.006 (0.076)
Education specialist	0.001 (0.034)	0 (0.021)	0 (0.018)
Masters	0.407 (0.491)	0.396 (0.489)	0.442 (0.497)
Nondegree	0.015 (0.12)	0.009 (0.095)	0.003 (0.059)
Years experience (total)	10.655 (10.04)	9.186 (9.32)	12.372 (8.853)
Years experience (district)	7.201 (7.841)	6.248 (7.485)	9.357 (7.536)
Salary	43878.79 (11970.023)	43765.152 (12517.188)	48351.581 (27839.638)
N	3,626	8,760	8,886

Note. VAS = value-added score. Variable means, standard deviations, and number of observations presented by step.

Table 4
Percentage of Teachers Using Measure in NMTEACH Rating, By Step

Step	VAS	Domains 2 and 3	Domains 1 and 4	Attendance	Surveys
1	0	100	92.6	77.2	42.9
2	100	100	93.9	82.6	58.3
3	100	100	95.9	83.0	67.1

Note. NMTEACH = New Mexico Educator Effectiveness System; VAS = value-added score.

When plotting the composite scores by step (see Figure 1), the differences in the step-specific score distributions become more apparent. While the mean composite scores for all three steps are within the bounds of the Level 3, or "Effective" category, Figure 1 makes clear that the distribution of summative scores for steps with higher VAS weights occupy a lower range of the

Table 5
Distribution of 2015–2016 NMTEACH Ratings, by Step

Step	Level 1, %	Level 2, %	Level 3, %	Level 4, %	Level 5, %	Average Points	*N*
Overall	5.50	23.20	42.50	24.90	3.90	131.90	21,272
1	0.80	7.30	51.70	33.70	6.50	142.07	3,626
2	2.40	24.10	47.50	23.80	2.10	132.14	8,760
3	10.50	28.70	33.70	22.40	4.60	127.51	8,886

Note. NMTEACH = New Mexico Educator Effectiveness System.

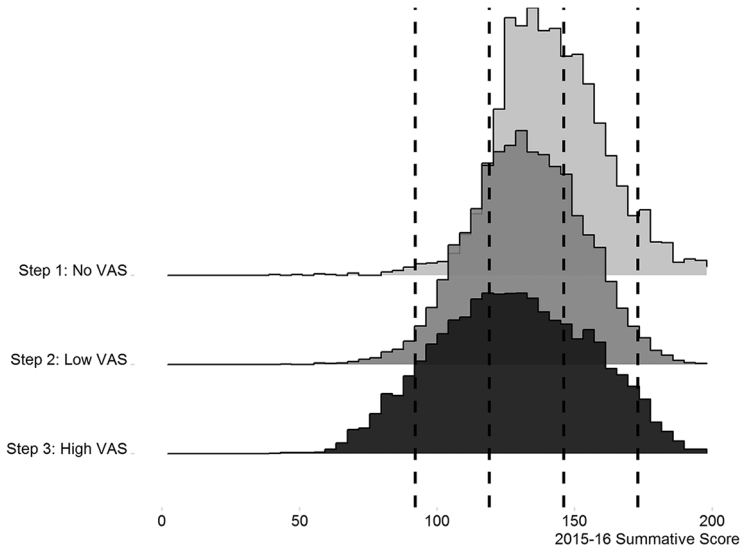

Figure 1. **Distribution of 2015–2016 NMTEACH summative scores, by step.**
Note. Vertical dotted lines indicate cut-points for NMTEACH ratings (1/2 = 92, 2/3 = 119, 3/4 = 146, 4/5 = 173). NMTEACH = New Mexico Educator Effectiveness System.

summative score scale. We later simulate changes to VAS weights to see its effect on rating consistency, though the differences in composites scores across steps also illustrate that such changes would also likely shift the distribution of composite scores (Steinberg & Kraft, 2017). Additionally, we see that the distribution of summative scores for Step 1 teachers is nonsymmetric, with bunching toward the higher end of the summative score range. This aligns with other literature (e.g., Grissom & Loeb, 2017) finding that administrator-issued classroom observations scores tend to be negatively skewed,

generally leading to teachers receiving higher scores on these measures than normally distributed value-added measures.

Methods

We focus specifically on estimating the within-year consistency of the summative NMTEACH ratings received by New Mexico teachers during the 2015–2016 school year. Summative ratings are distinct from composite scores, and throughout the article, we use the term *rating* to refer to the ordinal classifications (Level 1 to Level 5) received by teachers, whereas "score" is used to refer to teachers' 22–200 score, that is, subsequently used to assign their rating. Consistency is related to reliability and is sometimes referred to as the reliability of classifications (Lee, Hanson, & Brennan, 2002) but is a distinct concept in several ways. Steinberg and Kraft (2017) examine the similarity of the distribution of teacher ratings across multiple weighting schemes, calculating something akin to scale reliability, which is largely driven by the extent to which teacher performance is correlated across measures. Other researchers have focused on the year-to-year stability of teachers' composite scores (e.g., T. J. Kane et al., 2013). Estimates of year-to-year stability differ from our measure of within-year consistency in that "true" teacher ability is more likely to change across years. Empirical studies typically find that teacher effectiveness tends to increase as teachers accrue experience, with the rate of improvement sharpest early in a teacher's career (Papay & Kraft, 2015). Estimating the consistency of ratings based on year-to-year stability will confound these genuine changes in teacher effectiveness with inconsistency in the ratings themselves. Additionally, the various interventions based on teacher ratings (professional development for low-rated teachers, performance pay for high-rated teachers) imply that policymakers actively encourage teachers to improve over time and believe that changes to ratings are tied to genuine changes in their effectiveness. For this reason, we focus on estimating the consistency of composite ratings within-year where underlying teacher quality is more likely to be stable and any inconsistency we find is more likely to be the result of variance across factors (e.g., students, lessons, raters) that are typically considered measurement error.

Under current NMTEACH policy, all valid component scores for a teacher are used toward the calculation of their NMTEACH composite score and subsequent rating; there is no existing "second, independent assessment," as described by Cronbach et al. (1997), that could be used to estimate consistency. Therefore, estimating the consistency of ratings on repeated measurement requires the simulation of replicate scores. We approach this simulation in two ways. Our first method uses sample information and assumptions about the joint distribution of observed component scores to generate plausible replicate scores. We assign ratings to these replicate scores per NMTEACH rules and take the rate of agreement between two replicate

scores as an estimate of consistency. Our second approach eschews these parametric assumptions and adopts a bootstrap method, calculating replicate scores using randomly sampled subsets of teachers' actual scores. Our simulation method and consistency calculation are described below.

Parametric Simulation

First, we use the method adopted in Douglas and Mislevy (2010) and Martínez et al. (2016), hereafter the "parametric" method, where replicate measure scores are simulated using parameters estimated from the 2015–2016 NMTEACH data. This method has three distinct steps.

1. We simulate 10,000 true scores (representing 10,000 teachers) for all five NMTEACH measures from a disattenuated multivariate normal distribution (Bock and Peterson, 1975), obtained using sample means and covariances from a sample of teachers with nonmissing values for all five measures. Table 6 shows the observed (upper triangle) and disattenuated true score (lower triangle) correlations, with estimates of measure reliability on the diagonal.
2. For each teacher-by-measure combination, we draw two replicate measure scores from a normal distribution with mean equal to the true score for that teacher-by-measure and standard deviation equal to the standard error of measurement implied by the observed standard deviation and reliability for that measure.[2]
3. NMTEACH ratings are then calculated for both sets of replicate scores using NMTEACH business rules from 2015–2016. This simulation is done separately by NMTEACH step, resulting in a final sample size of 30,000 (10,000 × 3 Steps) simulated teachers.

This method depends greatly on estimates of measure reliability to estimate true scores and to generate replicate scores. Because different estimates of reliability could result in different estimates of consistency, we briefly describe our approach to estimating component measure reliability here. VAS reliability ($r = 0.64$) is estimated as the correlation of teachers' adjacent-year overall VAS scores.[3] Reliability for Domains 2 and 3 ($r = 0.62$) is estimated by identifying a subset of teachers who received two observations from distinct raters on two separate occasions and taking the correlation of these paired observations. Domains 1 and 4 reliability ($r = 0.96$) is based on a fully nested generalizability study decomposing observed score variance into occasion, teacher and observer components (Brennan, 2001; Mashburn, Meyer, Allen, & Pianta, 2014; Schweig, 2018).[4] Reliabilities for the student surveys ($r = 0.52$) were based on the intraclass correlation–type coefficients for scores averaged over multiple raters (Brennan, 2001; Shrout & Fleiss, 1979).

Nonparametric Bootstrap

Our bootstrap approach (Brennan & Wan, 2004) takes advantage of the fact that teachers often receive multiple observations over the course of the

Table 6
Observed and Disattenuated Correlations

	VAS	D23	D14	ATT	SVY
VAS	**0.636**	0.236	0.207	0.071	0.119
D23	0.348	**0.620**	0.810	0.046	0.245
D14	0.266	0.992	**0.955**	0.072	0.210
ATT	0.089	0.055	0.074	**1.000**	0.040
SVY	0.210	0.396	0.300	0.056	**0.515**

Note. VAS = value-added score. Table shows observed (upper triangle) and disattenuated (lower triangle) correlations. Correlations are disattenuated using the multivariate procedure described in Bock and Peterson (1975). Bolded values on diagonal are reliability estimates. VAS reliability is estimated as the correlation of teachers' adjacent-year overall VAS scores. Domains 2 and 3 reliability is estimated by identifying a subset of teachers who received two observations from distinct raters on two separate occasions and taking the correlation of these paired observations. Reliability for Domains 1 and 4 was estimated in a separate G-study (Schweig, 2018). Reliability for the student surveys were based on the intraclass correlation-type coefficients for scores averaged over multiple raters. Attendance reliability is assumed to be 1.

school year and/or receive separate value-added estimates across multiple subjects. Roughly 70% of New Mexico teachers in our sample are observed multiple times on at least one of the 10 observation items included in Domains 2 and 3 of the NMTEACH rubric and 60% of teachers with value-added estimates received VAS scores in two or more subjects, resulting in a sample size of 16,651 teachers for these simulations. Teachers who were included in the nonparametric sample were slightly lower performing but more experienced than colleagues excluded from the sample (see Appendix Table A1, in the online version of the journal, for full comparison table).

Official NMTEACH policy constructs an "overall" component score by simply averaging across a teacher's available scores during the policy-specified time frame (i.e., 1 year for observation scores, up to 3 years for VASs). To form replicates using our nonparametric approach, we calculate averages using a randomly sampled (with replacement) set of teachers' scores rather than all available scores using a four-step process:

1. For each teacher × measure combination, we randomly sample (with replacement) as many elements as were used for that teacher's actual 2015–2016 score.
2. Randomly sampled elements are averaged, within measure, to form a teacher's first replicate measure score.
3. Steps 1 and 2 are repeated to form a second set of replicate measure scores.
4. NMTEACH ratings are then calculated for both sets of replicate scores using NMTEACH business rules from 2015–2016.

Important, these samples may include multiple instances of the same score and no instances of others, whereas a teacher's actual component score uses a single instance of all available scores. Measures (or items within measures) on which teachers do not vary within year, such as attendance and Domain 1 and 4 scores, are taken as constant and "filled down" across replicates.[5]

Relative to the parametric method, these are several advantages of this approach. First, the bootstrap method does not require any assumptions about how teacher performance is jointly distributed across measures. Second, the bootstrap method does not require estimates of measure reliability, as it simply uses variation in scores across observation cycles and subjects as an empirical method for generating measurement error across replicates.

Consistency estimates obtained using the bootstrap method are likely positively biased due to the finite set of scores that we use to sample with replacement, particularly for the observation measures. In the nonparametric sample, the median teacher has six VA scores but only two sets of observation scores from which to resample. If a teacher has only two sets of scores for a given observation item, there is a 37.5% chance that both replicates will draw the same set of scores for that item. In contrast, because the parametric method draws replicates from an infinitely large set of values, the likelihood of exact matches for component scores across replicates is very slim, guaranteeing that each replicate will have some degree of deviation from the other. This issue also pertains to the sampling of teacher VAS but is mitigated because NMTEACH policy uses teacher VAS across all subjects over 3 years for a given teacher.

Comparing Simulated and Actual NMTEACH Scores

We assess the plausibility of our simulated NMTEACH scores by appraising the similarity of the simulated and empirical score distribution. In general, both the parametric and nonparametric closely overlap the observed score distributions and successfully recover the empirical means and covariances. However, the nonparametric method is more successful at recovering the skew of the Step 1 distribution and the kurtosis of the Step 3 distribution. This is illustrated in Figure 2, where the simulated score distribution is overlaid on the distribution of actual NMTEACH scores.

Calculating Consistency

Once NMTEACH ratings are assigned, we calculate, separately for the parametric and nonparametric replicates, a consistency statistic defined as the percentage of teachers who have the same 1–5 summative rating across both replicates. Maintaining the definition used elsewhere in the literature (Brennan & Wan, 2004; Douglas & Mislevy, 2010; Livingston & Lewis, 1995; Martínez et al., 2016), this definition of consistency refers to the agreement of two observed ratings that are measured with error.

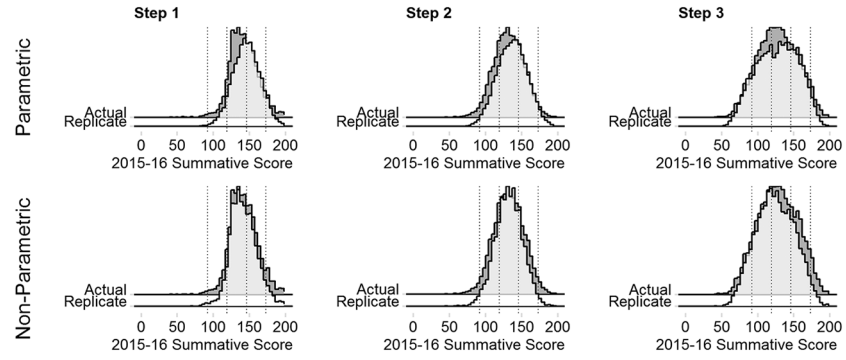

Figure 2. **Replicate and actual NMTEACH score distributions.**

Note. VAS = value-added score; NMTEACH = New Mexico Educator Effectiveness System. Clear figures in foreground are histograms of the replicate summative score distribution. Gray figures in background histograms of the actual 2015–2016 NMTEACH summative scores used to simulate the replicate scores. The parametric method uses a sample of teachers who have scores on all five measures. The nonparametric method uses a sample of teachers of have multiple scores on the Domains 2 and 3 observation and/or VAS measures, depending on step.

Policy Conditions

To answer the second research question, we conduct policy simulations to examine the effects of three potential mechanisms for improving rating consistency. The goal of these simulations is to illustrate the gains (or losses) to consistency that policymakers might expect to see following changes to (1) measure reliability, (2) measure weights, and (3) the number and location of rating cut-points.

First, we examine the relationship between measure reliability and composite rating consistency. Intuitively, increases in measure reliability will likely result in increases to composite rating consistency. Unlike the other policy conditions that we explore through simulation, changes to measure reliability cannot be directly imposed. Instead, policymakers will need to modify other evaluation policies, which will improve measure reliability indirectly. These policies might include increasing the number of years of student test score data that are included in a teacher's VAS calculation, increasing the number of required classroom observations, or imposing a minimum number of students to be assigned to a teacher evaluated using student-level scores (i.e., VAS, student surveys). These policy modifications will result in increases to the reliability of the relevant component measures, and subsequently, the consistency of the composite rating. We explore this relationship between measure reliability and rating consistency empirically

by repeating our parametric simulations and varying the inputted reliability of (1) the Domain 2 and 3 observation scores and (2) VAS with each iteration. This method allows us to see the full range in the relationships among VAS reliability, observation score reliability and rating consistency. Observation score and VAS reliabilities higher than what is used in our parametric simulations ($r = 0.62$ and $r = 0.64$, respectively) are included to suggest improvements to consistency that may come as result to implementation of policies that increasing score reliability. Conversely, because our empirical estimates of both VAS and observation score reliabilities are higher than is typically reported in the literature, simulations using lower estimates of score reliability provide important sensitivity analyses and can be useful for applying our findings to other evaluation systems.

Second, we consider the effect of changing the weights of individual measures on consistency. Changes to the weights assigned to component measures are typically discussed as a mechanism to change the distribution of the final composite scores (Steinberg & Kraft, 2017). However, such changes will also affect the consistency of resulting composite rating. If weights are shifted toward more reliable components and away from less reliable ones, we would expect rating consistency to increase as well. Similar to our investigation of measure reliability, we can plot the relationship between measure weight and rating consistency by running successive simulations, changing the weights assigned to each measure with each iteration. Specifically, we begin our simulation using the Step 1 weights and repeat the simulation 100 times, each time assigning an additional percentage point to teachers' VAS and distributing the remaining weight proportionally among the other measures. We run these iterative simulations using both parametric and nonparametric methods. We restrict this analysis to Steps 2 and 3 teachers, since changes to VAS weight will have no effect on Step 1 teachers who do not have VAS scores.

The third mechanism for affecting rating consistency is adjusting the location and number of rating cut-points. Cut-point placement does not affect the underlying amount of measurement error for a given score but can exacerbate the consequences of measurement error. Assuming homoskedastic measurement error, observations closer to rating cut-points are more likely to be inconsistently rated. By extension, the sample-wide consistency rate will be lower if rating cut-points are located in denser regions of the underlying summative score distribution. We illustrate the importance of cut-point location by examining how consistency around a single cut-point varies as the cut-point location on the 22–200 summative score scale changes with each iteration of the simulation. Additionally, if proximity to a cut-point decreases consistency, removing a cut-point altogether, thus condensing the number of rating categories, will increase it. While NMTEACH uses a five-level rating system, other evaluation systems across the country typically use only three or four different levels of ratings (Kraft & Gilmour, 2017).

Table 7
Consistency Estimates Using Parametric Replication Method

Step	Exact, %	±1, %	±2, %	Corr.	SEM
Overall	60	97	100	0.79	12.20
Step 1	72	100	100	0.88	6.44
Step 2	62	99	100	0.79	9.34
Step 3	45	92	100	0.70	15.73

Note. This table presents estimates of rating consistency using the parametric simulation method. Consistency compares ratings calculated from observations' replicate Score 1 and replicate Score 2. "Exact," "±1," and "±2" show the percentage of cases whose ratings match exactly within 1 level and within 2 level, respectively. "Corr" is the correlation between the observations' underlying summative scores. SEM is the implied standard error of measurement, calculated as $SEM = SD^* \sqrt{(1 - Corr)}$, where *SD* is the observed standard deviation of replicate score 1 and *Corr* is used as an estimate of measure reliability.

Therefore, we also estimate the consistency of a hypothetical three category rating system. We do this by collapsing the five-level rating NMTEACH system to a three-category system in two ways: (1) a "tail-heavy" system that collapses Levels 1 and 2 and 4 and 5, leaving Level 3 unchanged and (2) a "center-heavy" distribution that collapses Levels 3, 4, and 5, leaving Levels 1 and 2 unchanged.

Results

Parametric Results

Table 7 shows estimates of the overall and step-specific consistency rates of NMTEACH ratings. In addition to rates of rating consistency, we also provide the correlation between teachers' (22–200) replicated summative scores, which can be interpreted as a measure of the summative scores' reliability, and the implied standard error of measurement of the summative score. Using parametric simulation, we estimate that 60% of teachers would obtain the same NMTEACH rating on remeasurement. The degree of deviation is generally limited to within one rating level, as 97% of teachers received replicate ratings that were at least ±1 level of each other. In addition, we find that as steps increase in the weight given to VAS, measures of consistency drop accordingly. The consistency rate of Step 1 teachers (72%) is roughly 30 percentage points higher than that of Step 3 teachers (45%). Encouragingly, 92% of Step 3 teachers still receive replicate ratings within 1 level of each other.

Appendix Table A2 (available in the online version of the journal) provides a complete transition matrix disaggregated by rating level for teachers' ratings across replicate scores obtained using the parametric method. Values

Table 8
Consistency Estimates Using Nonparametric Replication Method

Step	Exact, %	±1, %	± 2, %	Corr.	SEM
Overall	75	98	100	0.86	8.63
Step 1	95	100	100	0.99	1.50
Step 2	80	100	100	0.90	5.81
Step 3	63	96	99	0.81	11.67

Note. This table presents estimates of rating consistency using the nonparametric simulation method. Consistency compares ratings calculated from observations' replicate Score 1 and replicate Score 2. "Exact," "±1," and "±2" show the percentage of cases whose ratings match exactly within 1 level and within 2 level, respectively. "Corr" is the correlation between the observations' underlying summative scores. SEM is the implied standard error of measurement, calculated as $SEM = SD^* \sqrt{(1 - Corr)}$, where *SD* is the observed standard deviation of replicate score 1 and *Corr* is used as an estimate of measure reliability.

in the transition matrix are row percentages indicating the percentage of observations receiving the column level (Replicate 2), conditional on receiving the row level (Replicate 1).

Nonparametric Results

Using the nonparametric replicates, we see similar general patterns as those obtained from the parametric replicates, with rates of consistency that are 15–20 percentage points higher when using the nonparametric replication method (see Table 8). As previously explained, we believe these nonparametric consistency rates may be positively biased as result of the small person-by-measure samples we are bootstrapping from. Consistent with the pattern observed with the parametric replicates, increases in the weight given to VAS are shown to have an adverse effect on rating consistency. Appendix Table A3 (available in the online version of the journal) provides transition matrices, disaggregated by rating level for teachers' ratings across replicate scores obtained using the nonparametric method.

Policy Simulations

Next, we conducted policy simulations to examine how rating consistency would be affected by changes to three aspects of the NMTEACH system: (1) measure reliability, (2) measure weights, and (3) the number and location of rating cut-points.

Changing Observation Score Reliability

First, we examine how rating consistency changes in response to changes to reliability of Domain 2 and Domain 3 observation scores.

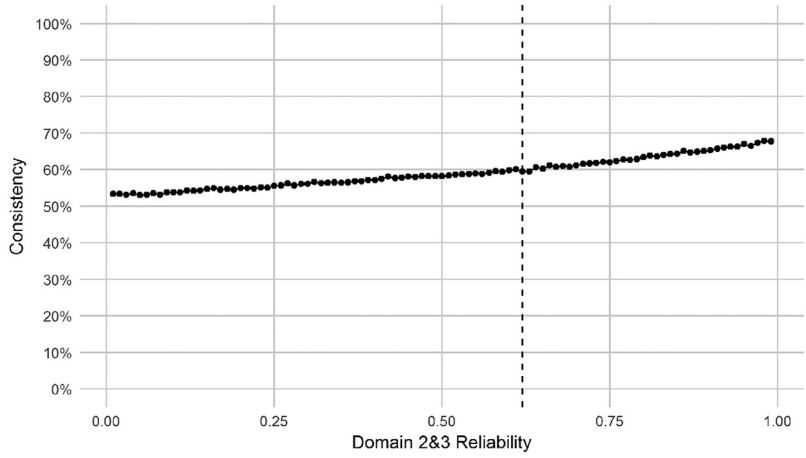

Figure 3. **Changes to rating consistency for Step 1–3 teachers, by observation score reliability.**

Note. Vertical dotted line indicates estimate of observation score reliability used in parametric simulation, .62.

Figure 3 shows replicate consistency estimates across repeated parametric simulations for hypothetical teachers in all 3 steps, with each iteration of the simulation using a different reliability estimate for observation scores. The vertical dashed line in the figure indicates the reliability estimate used in our parametric estimates, which equals .62.

As expected, increases to the reliability of the observation score results in increases to rating consistency. However, there are relatively small declines in consistency as the reliability of the observation scores decreases. If observation score reliabilities are decreased to around .50, for example, the overall consistency rate drops by only 2 percentage points relative to the results reported in Table 7 (60% to 58%). This is potentially an important result for policymakers, as the .50 reliability is close to observation score reliabilities reported elsewhere, including those reported as a part of the MET project (Cantrell & Kane, 2013).

Changing VAS Reliability

Next, we examine how rating consistency changes in response to changes to VAS reliability. Figure 4 shows replicate consistency estimates across repeated parametric simulations for Step 2 and 3 hypothetical teachers, with each iteration of the simulation using a different reliability estimate for VAS. The vertical dashed line in the figure indicates the VAS reliability

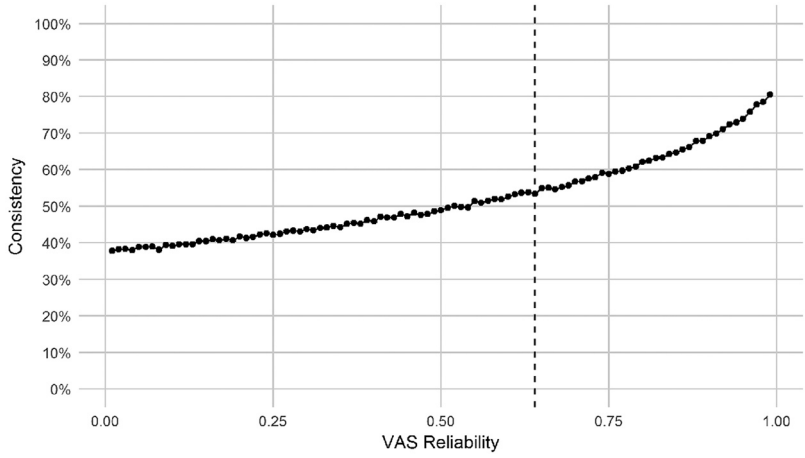

Figure 4. **Changes to rating consistency for Steps 2/3 teachers, by VAS reliability.**
Note. VAS = value-added score. Vertical dotted line indicates estimate of VAS reliability used in parametric simulation, .64.

estimate used in our parametric estimates, .64. Step 1 hypothetical teachers are omitted because their rating consistency is unaffected by changes to VAS reliability.

As expected, increases to the reliability of VAS also result in increases to rating consistency. Simulations suggest that for value-added reliability estimates that are at the lower end of those typically encountered in practice ($r = 0.20$), consistency rates for Step 2 and 3 teachers drop by roughly 11 percentage points relative to the results reported in Table 7. Our simulation suggests that the returns to rating consistency are nonlinear, with the marginal effect of VAS reliability increasing as VAS approaches perfect reliability.

Changing VAS Weight

To plot the relationship between measure weight and value-added reliability, we again use repeated simulation, this time, iterating the weight given to VAS. Weights for non-VAS measures are assigned such that they maintain the ratio used in Step 1, that is, the Domain 2 and 3 scores, Domain 1 and 4 scores, teacher attendance, and student surveys always occupy 50%, 40%, 5%, and 5% of the non-VAS weight, respectively. The consistency rates from both the parametric and nonparametric methods are plotted in Figure 5.

Similar to the effect of changes on VAS reliability, increases to VAS weight exhibit the expected negative effect on rating consistency, but the

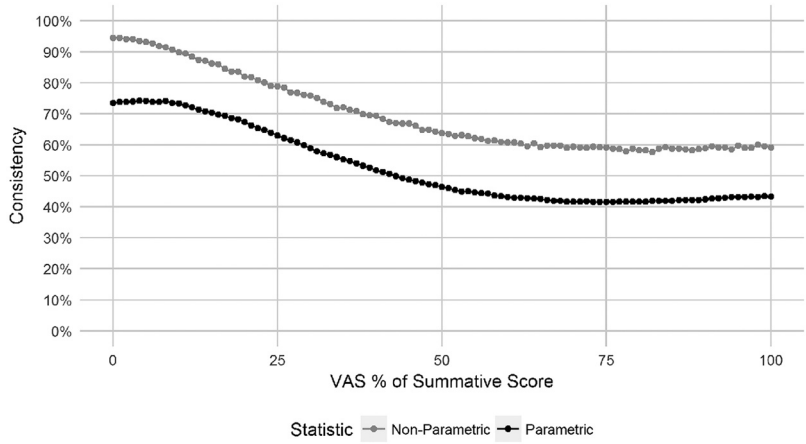

Figure 5. **Changes to rating consistency for Steps 2/3 teachers, by VAS weight.**
Note. VAS = value-added score. Step 1 weights are used for all measures in the "VAS % = 0" iteration. For each additional iteration of the simulation, non-VAS measures are assigned the remaining weight not assigned to VAS, maintaining the same ratio found in Step 1, for example, the Domain 2 and 3 score will always occupy 50% of the non-VAS weight and the Domain 1 and 4 score will always occupy 40% of the non-VAS weight.

relationship between these quantities also appears nonlinear: Increases in VAS weight from 0% to 60% have a roughly linear negative relationship with composite reliability but appear to have no effect on reliability after the 60% mark. This plateau is unlikely to affect policy decisions, since the range of VAS weights currently used in most multiple measure systems are within 0% to 50%. Within the observed range across teacher evaluation systems, our findings suggest that decisions to adjust the weights of VASs have a pronounced effect on rating consistency.

Changing NMTEACH Cut-Points

Last, we describe the role of cut-point location on the consistency of composite ratings by incrementing the location of a single cut-point along the 200-point NMTEACH summative score scale. The consistency statistic in this simulation is the percentage of teachers who are consistently rated above or below the cut-point location for that given simulation. This statistic is analogous to calculating a consistency rate for a two-category system, which can be a useful result for systems that have multilevel composite ratings but attach consequences to only one of these ratings. Figure 6 shows the consistency rates as a function of cut-point location, disaggregated by method and step.

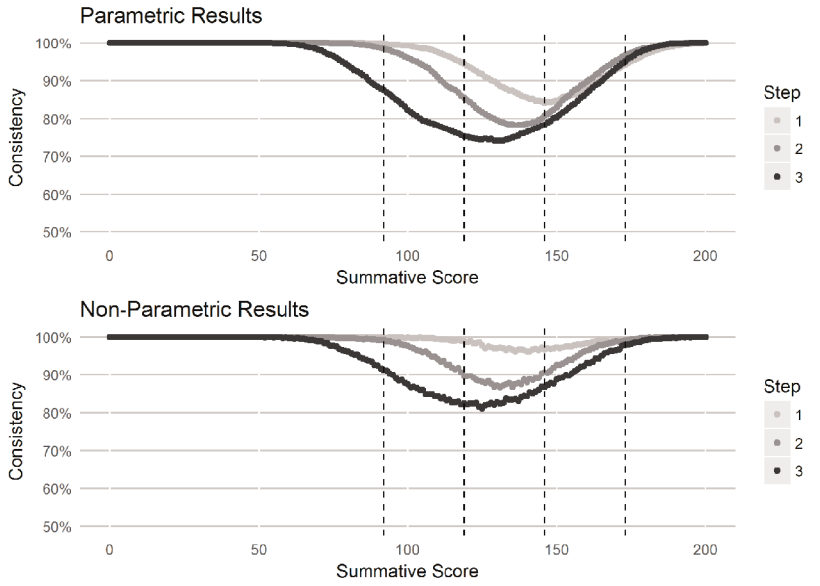

Figure 6. **Changes to consistency of teacher ratings, by cut-point location.**

Note. Figure shows the consistency rate for a two-category rating system (i.e., below/above cut-point), iterating the cut-point location from 0 to 200. Vertical dotted lines indicate cut-points for NMTEACH ratings (1/2 = 92, 2/3 = 119, 3/4 = 146, 4/5 = 173). NMTEACH = New Mexico Educator Effectiveness System.

We find that the cut-point location at which composites are least reliable vary by step. The consistency curves are essentially mirror images of the composite score distributions, where the "valleys" of the former are located at the "peaks" of the latter. The valley of the consistency curves, that is, the cut-points of lowest consistency, for Steps 1 and 3 teachers are at the 3/4 and 2/3 boundaries, respectively, with the location of the Step 2 valley split in-between.

In this same vein, we examine how consistency rates would be affected by reducing the number of rating categories from five (the current NMTEACH system) to three. We construct both a "tail-heavy" three category system, which combines Levels 1 and 2 and Levels 4 and 5, and a "center-heavy" three category system, which combines Levels 2, 3, and 4. Cut-points for the tail-heavy distribution will be closer to the center of the distribution than those for the center-heavy distribution. Therefore, examining differences between the consistency rates for the tail-heavy and center-heavy systems allow us to examine the effects of category reduction as well as cut-point location on rating consistency. Parametric results for the tail-heavy

Table 9
Three Category Parametric Results (Tail-Heavy)

Step	Exact, %	±1, %	±2, %	Corr.	SEM
Overall	68	98	100	0.79	11.29
Step 1	79	100	100	0.88	6.48
Step 2	66	100	100	0.80	9.27
Step 3	59	95	100	0.69	15.97

Note. This table presents estimates of rating consistency for a hypothetical "tail-heavy" three category rating system using the parametric simulation method. Tail-heavy categories were created from the original 5 categories ratings by combining Levels 1 and 2, leaving Level 3 intact, and combining Levels 4 and 5. Consistency compares ratings calculated from observations' replicate score 1 and replicate score 2. "Exact," "±1," and "±2" show the percentage of cases whose ratings match exactly within 1 level, and within 2 level, respectively. "Corr" is the correlation between the observations' underlying summative scores. SEM is the implied standard error of measurement, calculated as $SEM = SD * \sqrt{(1 - Corr)}$, where *SD* is the observed standard deviation of replicate score 1 and *Corr* is used as an estimate of measure reliability.

and center-heavy three categories are in Tables 9 and 10, respectively. Nonparametric results are provided in Appendix Tables A4 and A5 (available in the online version of the journal).

Relative to the overall results, consistency rates are higher for a hypothetical three category system than for a five-category system. However, the reliability statistics measured using the NMTEACH summative scores, rather than ratings, do not change. This is consistent with the fact that changing or removing cut-points does not affect the amount of measurement error in the underlying summative scores, only the risk of being inconsistently classified. While the overall consistency rates for both the tail-heavy (Parametric: 68%) and center-heavy (90%) three category ratings are improvements from the consistency of the original five category ratings (60%), cut-point location clearly matters for rating consistency. Consistent with our "single cut-point" analysis presented in Figure 5, steps with ratings closer to the densest regions of the underlying score distribution (the tail-heavy system) will be more prone to inconsistency than those with cut-points in the fringes (the center-heavy system).

Discussion

This analysis estimates the consistency of composite ratings of teacher effectiveness issued from an at-scale high-stakes teacher evaluation system in New Mexico. Using two methods to simulate a reevaluation of teachers during the 2015–2016 academic year, we find that between 25% (nonparametric) and 40% (parametric) New Mexico teachers would be expected to

Table 10
Three Category Parametric Results (Center-Heavy)

Step	Exact	±1	±2	Corr.	SEM
Overall	90%	100%	100%	0.79	11.29
Step 1	94%	100%	100%	0.88	6.48
Step 2	95%	100%	100%	0.80	9.27
Step 3	82%	100%	100%	0.69	15.97

Note. This table presents estimates of rating consistency for a hypothetical "center-heavy" three category rating system using the parametric simulation method. Center-heavy categories were created from the original 5 categories ratings by leaving Level 1 intact, combining Levels 2, 3, and 4 and leaving Level 5 intact. Consistency compares ratings calculated from observations' replicate score 1 and replicate score 2. "Exact," "±1," and "±2" show the percentage of cases whose ratings match exactly within 1 level, and within 2 level, respectively. "Corr" is the correlation between the observations' underlying summative scores. SEM is the implied standard error of measurement, calculated as $SEM = SD * \sqrt{(1 - Corr)}$, where *SD* is the observed standard deviation of replicate score 1 and *Corr* is used as an estimate of measure reliability.

receive a different rating if they were reevaluated. Teachers whose composite scores rely more heavily on VASs are more likely to be rated differently according to our findings from both simulation methods. Consistency among Step 1 teachers, whose composite scores do not incorporate VAS, is roughly 30 percentage points higher than consistency rates for Step 3 teachers, whose composite scores are 50% determined by VAS.

Perhaps our most striking finding is the decreases in rating consistency associated with increases in VAS weight. However, policymakers looking to act on our estimates should consider a number of important caveats. First, while the design of NMTEACH certainly shares many features with other multiple measure evaluation systems that have emerged in the Race to the Top era, the New Mexico system is distinct in several ways, including its use of teacher attendance and survey measures and the roughly normal distribution of its final summative ratings. Additionally, even for commonly used measures, such as observation scores and value-added, the consistency estimates we obtain in this article are specific to the properties of these measures (e.g., reliabilities, correlations with other measures) as implemented in NMTEACH. Our finding that rating consistency is improved when shifting weight away from VAS and toward observation scores is dependent on the fact that, using the NMTEACH data, we estimate that non-VAS components are relatively more reliable than VAS. In systems where observation scores are substantially less reliable than VASs, giving more weight to observation score components, intuitively, may have a negative effect on rating consistency.

Taken on their own, the consistency estimates that we obtain in this study offer a cautionary tale that recommends against using composite

ratings, derived from systems with properties similar to NMTEACH, to make high-stakes inferences regarding the effectiveness of individual teachers. However, it is important to remember that these results *should not* be considered in isolation. The analyses presented here focus exclusively on consistency and do not address other important properties of a valid and useful evaluation system, including (most important) accuracy, cost, and coverage. For example, prior teacher evaluation systems where virtually all teachers were deemed "satisfactory" on the basis of informal observations produced remarkably consistent but highly inaccurate evaluation scores for most teachers. Scores produced by NMTEACH and other multiple measure systems are almost certain to be less consistent than scores produced under these prior systems. However, this reduction in consistency comes from the incorporation of measures that are more accurate and provide more actionable feedback based on measures that are more closely aligned to student outcomes (Chetty et al., 2014; T. J. Kane et al., 2011).

In many ways, this is similar to what statisticians refer to as a "bias-variance tradeoff": policy changes that improve rating consistency may also adversely affect these other desirable properties. We demonstrate that by shifting weight away from less reliable measures, such as value-added in the context of NMTEACH, the consistency of summative ratings will increase. However, policymakers will ultimately need to consider evidence of score consistency in combination with evidence of score accuracy as well other sources of evidence about score validity, and with cost and feasibility considerations.

Evidence regarding the predictive power of observation scores for long-run student outcomes are not yet available given the recent adoption of formal observations. However, recent studies finding evidence of principal-driven (Grissom & Loeb, 2017) and student characteristic-driven (Steinberg & Garrett, 2016) bias raise concerns about the quality of information captured by observational measures. Shifting more weight toward classroom observation scores will generally produce a more consistent signal, but this signal may not be an accurate indication of the teachers who are successful at promoting student success.

Based on our "policy simulations," one clear implication of our results is that if consistency in teacher ratings is valued, efforts should be made to examine and improve the reliability of constituent measures, including both the observation-based measures and value-added measures used in teacher evaluation systems without changing the underlying model used to combine measures. This may include conducting more frequent teacher observations, having observations conducted by multiple raters, or increase the number of years of VASs averaged to form composites. In particular, any systems using 1- or 2-year averages of VAS should consider requiring teachers to have at least 3 years of estimates prior to using these scores for accountability. While such a policy precludes early career teachers from having value-added included in their evaluation score, this may be a tradeoff

that policymakers are willing to make. Similarly, increasing the number of minimum student test scores or survey responses will subsequently result in increases to the reliability of the VAS and survey components, respectively, but will come at the cost of reducing the coverage of teachers with valid measures. Additionally, we demonstrate how reducing the number of categories or shifting the rating cut-points out toward the tails of the summative score distribution can improve the consistency of their ratings. Since additional categories increase the risk of rating inconsistency, policymakers should consider keeping only categories that are attached to specific interventions and of summative or formative importance.

The weight given to each measure and the number and location of the cut-points should depend on how policymakers and practitioners balance the need for validity with the tolerance for misclassification. The need for increased consistency, given trade-offs to other properties, will differ by system, varying according to how composite scores and ratings are used (Messick, 1995). In general, as the penalties for poor performance increase, stakeholders' tolerance for misclassification should decrease. Our analysis does not provide guidance as to whether systems *should* pursue the various mechanisms we outline. Rather, we offer our estimates of NMTEACH consistency as a "business-as-usual" baseline for policymakers operating high-stakes teacher evaluation systems. If policymakers opt to modify evaluation systems, our various policy simulations demonstrate the potential effects of these changes on rating consistency. Ultimately, the structure of these systems and any changes to them based on our analysis should depend on the aims of a teacher evaluation system as identified by its stakeholders.

Notes

This research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160223. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. Supplemental material is available for this article in the online version of the journal.

[1]Readers interested in a more comprehensive synopsis of the value-added literature are encouraged to refer to Braun (2005), Koedel et al. (2015), and McCaffrey, Lockwood, Koretz, and Hamilton (2004).

[2]The standard error of measurement for any given measure is calculated as $SD^* \sqrt{(1 - r)}$, which is the measure's sample standard deviation multiplied by the square root of one minus the estimate of reliability for that measure.

[3]Existing estimates of value-added reliability generally range from .2 to .7 (McCaffrey, Sass, Lockwood, & Mihaly, 2009), with our estimate of VAS reliability on the higher end of this range. However, the published estimates range widely with regard to both (1) model specification and (2) the method used to estimate reliability. Estimates in the lower end of this range tend to be estimated using the adjacent-year correlations of single-year, single-subject value-added estimates. As a multiyear, multisubject average, it is reasonable that the reliability of the "overall" VAS used in NMTEACH is on the higher end of this spectrum.

[4]We are unable to estimate Domain 1 and 4 reliability in the same way as Domain 2 and 3 due to the extremely small sample (N ~20) of teachers receiving Domain 1 and 4 scores from different raters.

[5]The "filling down" used in the nonparametric simulation, in essence, assumes that these components have perfect reliability and is a source of positive bias in our consistency estimates. However, the size of this bias is very small, given that the filled down components already have relatively high-estimated reliabilities and/or are assigned a very small component weight.

References

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, *44*, 448–452.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (Working Paper No. 23478). Cambridge, MA: National Bureau of Economic Research.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper No. 278). Retrieved from http://www.epi.org/files/page/-/pdf/bp278.pdf

Balch, R. (2012). *The validation of a study survey on teacher practice*. Retrieved from https://www.researchgate.net/publication/265225866_THE_VALIDATION_OF_A_STUDENT_SURVEY_ON_TEACHER_PRACTICE

Bill & Melinda Gates Foundation. (2010). *Working with teachers to develop fair and reliable measures of effective teaching*. Retrieved from https://docs.gatesfoundation.org/Documents/met-framing-paper.pdf

Bock, R. D., & Petersen, A. C. (1975). A multivariate correction for attenuation. *Biometrika*, *62*, 673–678.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*, 3–15.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/PICVAM.pdf

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295–317.

Brennan, R. L., & Wan, L. (2004, June). *A bootstrap procedure for estimating decision consistency for single-administration complex assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2572&filename=MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Cherng, H.-Y. S., & Halpin, P. F. (2016). The importance of minority teachers: Student perceptions of minority versus white teachers. *Educational Researcher*, *45*, 407–420.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, *22*(2), 32–41.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*, 2633–2679.

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education*, *54*(1), 3–5.

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education Finance and Policy*, *8*, 418–434.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399.

Cullen, J. B., Koedel, C., & Parsons, E. (2016). *The compositional effect of rigorous teacher evaluation on workforce quality*. Cambridge, MA: National Bureau of Economic Research.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*, 267–297.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, *35*, 280–306.

English, D., Burniske, J., Meibaum, D., & Lachlan-Haché, L. (2016). *Using student surveys as a measure of teaching effectiveness*. Washington, DC: American Institutes for Research.

Ferguson, R. F. (2010). *Student perceptions of teaching effectiveness* (Discussion Brief). Cambridge, MA: National Center for Teacher Effectiveness and the Achievement Gap Initiative.

Goe, L., Holdheide, L. R., & Miller, T. (2011). *A practical guide to designing comprehensive teacher evaluation systems: A tool to assist in the development of teacher evaluation systems*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.lauragoe.com/LauraGoe/practicalGuideEvalSystems.pdf

Goldhader, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*, 199–208.

Grissom, J. A., & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy*, *12*, 369–395.

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

Hill, H. C., Umland, K. L., Litke, E., & Kapitula, L. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education*, *118*, 489–519.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2520&filename=MET_Reliability-of-Classroom-Observations_Research-Paper.pdf

Jackson, C., & Mackler, K. (2016). *Assessing effectiveness: How urban teachers evaluates its new teachers*. Baltimore, MD: Urban Teachers.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement*, *48*(1), 12–30.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, *17*, 221–240.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Retrieved from http://k12education.gatesfoundation.org/down

load/?Num=2676&filename=MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*, 587–613.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*, 234–249.

Kyriakides, L. (2005). Extending the comprehensive model of educational effectiveness by an empirical investigation. *School Effectiveness and School Improvement*, *16*, 103–152.

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412–432.

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, *78*, 75–95.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, *38*, 738–756.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*, 400–422.

McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*, 572–606.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 5D*, 741–749.

Mihaly, K., & McCaffrey, D. F. (2014). Grade level variation in observational measures of teacher effectiveness. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 9–49). San Francisco, CA: Jossey-Bass.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

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

Panorama. (2015). *Validity brief: Panorama student survey*. Retrieved from https://panorama-www.s3.amazonaws.com/files/panorama-student-survey/validity-brief.pdf

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S., & La Paro, K. M. (2006). *Classroom Assessment Scoring System (CLASS) manual: Middle/secondary version pilot*. Charlottesville: Curry School of Education, University of Virginia.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET Project*. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Schweig, J. D. (2018). *Pilot to policy: Reconsidering the reliability of observation-based ratings from New Mexico's statewide teacher evaluation system*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems*. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/technical_reports/TR917.html

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*, 340–359.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*, 293–317.

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*, 378–396.

Strunk, K. O., Weinstein, T. L., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives*, *22*, 1–41.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*, 3628–3651.

U.S. Department of Education. (2010). *Race to the top program guidance and frequently asked questions*. Retrieved from https://www2.ed.gov/programs/racetothetop/faq.pdf

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal*, *53*, 1834–1868.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Retrieved from https://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wise, E. A., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1985). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/content/dam/rand/pubs/reports/2006/R3139.pdf

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education*, *12*, 236–247.

Manuscript received May 23, 2018
Final revision received January 9, 2019
Accepted February 26, 2019