Cross-Year Stability in Measures of Teachers and Teaching

Heather C. Hill
Mark Chin
Harvard Graduate School of Education

In recent years, more stringent teacher evaluation requirements have focused attention on new metrics for assessing teacher and teaching quality[1]. One important issue is the degree of cross-year stability for these key metrics. Many assume teacher quality is a relatively stable underlying trait; outside of trends over time that might result from professional development, grade level or curricular change, teachers tend to teach the same material in a similar manner year after year, often with the same level of content knowledge and other supportive resources. If scores on contemporary indicators of teacher quality prove to vary substantially from year to year without explanation, stakeholders – including teachers themselves – may call into question the validity of conclusions about teacher quality based on those scores. In fact, substantial variability seems to be the case: existing evidence on the stability of many teacher measures suggests that cross-year stability is low to moderate. While the process-product research suggested cross-year correlations in observed teacher behaviors are generally above 0.5 (Brophy, Coulter, Crawford, Evertson, & King, 1975), Polikoff (2013) shows that the cross-year stability of observational measures from the more recent Measuring Effective Teaching (MET) study (Kane & Staiger, 2012) range from 0.3 to 0.4. MET student reports of classroom quality, aggregated for use at the teacher level, showed similar stability. Value-added scores, an important component of many teacher evaluation systems, appear to have cross-year correlations between 0.2 and 0.5 (Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009). These findings imply that teacher evaluation scores may shift markedly between years. Replicating and extending these findings in a sample of fourth and fifth grade teachers with a variety of teacher accountability metrics is one focus of this paper.

Another focus is the extent to which these shifts can be explained by changes in classroom composition, teacher learning, or other differences between years of instruction. Though some differences between adjacent-year scores observed in research can be attributed to measurement error, other factors may contribute to differences as well. For instance, some portion of cross-year differences may be responsive to the students in the classroom or the provision of professional development and/or other resources for teaching. If substantiated empirically, this would suggest that at least some component of the cross-year deviations in teacher scores reflect real changes in classroom conditions as opposed to measurement error.

Of particular interest in the search for explanations for cross-year differences is the extent to which teachers' reports of student "quality" correlate with both observers' estimates of classroom quality and teachers' own value-added scores. If teachers can predict changes in their own value-added scores based on estimates of average student ability, it both validates and calls into question the use of those scores for classroom accountability.

To gain insight into these issues, we use data from fourth and fifth-grade teachers of mathematics and their students. Teachers were recruited from four urban districts and followed over three

[1] For simplicity, we will refer to both as "teacher quality" unless we are talking about one specifically.

years. Data included items and constructs from teacher and student surveys, student administrative and test score data, and digital recordings of up to three lessons per year scored using both mathematics-specific and general pedagogical observational instruments. To investigate the cross-year stability of measures of teacher quality, we identify the percent of total variance in teacher scores on these measures attributable to solely teachers, after controlling for the year of data collection. We decompose teacher-level scores over either two (value-added) or three (observational and student survey metrics) adjacent school years, and then explore whether cross-year differences[2] can be explained by changes in classroom composition, teacher resources, or other factors.

Literature review

In this section, we review existing research on cross-year stability in teacher quality metrics.

Value-added metrics. The stability of measures aggregated from student test scores has been of concern since the early 1970s, when researchers began to identify more and less effective teachers based on their students' gains on basic skills assessments. In these studies, often referred to collectively as the process-product literature, correlations between adjacent-year gains in student test scores measured ranged from 0.2 to 0.4 (Brophy, 1973; Brophy et al., 1975; Good & Grouws, 1975). However, teacher-level scores in this literature often represented simple gains (post-test differenced from pre-test) rather than true value-added models, which parameterize the calculation of scores and often control for student and classroom characteristics.

More recently, many scholars have examined cross-year stability in teacher value-added scores. Some have done so by calculating contingency tables describing adjacent-year value-added ranks. Koedel and Betts (2007), for instance, found in San Diego data that only 20 to 35 percent of teachers remained in the same performance quintile in consecutive years; 13% moved from the first to last quintile, or vice versa. Using Chicago data, Aaronson, Barrow, and Sander (2007) find 26 to 57% of teachers remained in quartile across years and that 18% of teachers changed from the top to the bottom (or vice versa) quartile. Ballou (2005) presented results from TVAAS that were also consistent with the studies described above.

Several other scholars calculated correlations or average correlations across years in order to estimate the degree of stability in teacher value-added scores. McCaffrey et al. (2009) used panel data from four Florida districts and found most cross-year correlations in the range of 0.2 to 0.3. Goldhaber and Hansen (2013) used ten years of panel data from North Carolina, finding that the average cross-year correlation between teacher scores was 0.55; interestingly, correlating a three-year average with a subsequent three-year average improved the average correlation to 0.65. Goldhaber and Hansen also conducted a variance decomposition of this panel data, finding that 34% of the variance in value-added scores is "dynamic", representing a long-term time trend, suggesting that other variables, such as teacher professional development or changes in instruction, might affect teacher performance on the measure. Finally, the MET study found a similar cross-year correlation of 0.2 for English Language Arts value-added, and roughly 0.5 for math value-added (Kane & Staiger, 2012). Disattenuation for measurement error can raise

[2] For simplicity, we refer to the target of our analyses as cross-year 'differences'. Operationally, instead of predicting actual differences in measure scores from one year to the next, we predict current year scores controlling for prior year scores on the same measure.

reported cross-year correlations (McCaffrey et al., 2009); however, the extent to which either disattenuation or three-year averages would prove useful to most districts is unknown, as most high-stakes personnel decisions are made in the first two years of a teacher's career. Furthermore, correcting correlations for measurement error does not necessarily adjust the individual scores of teachers for error.

Observational metrics. Within the process-product studies of the 1970s, scholars also worried about the stability of the "process" side of the equation – teachers' classroom behaviors – used to predict aggregated student learning gains. Some scholars (Brophy et al., 1975; Marshall, Green, Hartsough, & Lawrence, 1977) examined stability in observation metrics across lessons within a given school year, noting that for many areas, considerable instability across time of day, subject matter, and lessons observed existed. This led to the application of generalizability theory (Shavelson & Webb, 1991) to attempt to recover within-year estimates of the stability of teacher behavior. Studies applying generalizability theory to modern instruments have estimated that between 13 to 40 percent of the variance in observation scores lies at the teacher level (Bell et al., 2012; Hill, Charalambous, & Kraft, 2012), leading many to recommend that such scores be based on multiple lessons assessed by multiple raters in order to improve overall reliability.

Cross-year stability in teacher metrics has been less often investigated. Brophy and colleagues (1975) estimated cross-year stability correlations for a set of classroom indicators observed four times in the first year and 14 times in the second. Correlations were in the 0.5 to 0.7 range for items capturing negative and positive teacher affect, clarity of the presentation, and teacher-initiated problem-solving. Polikoff (2013) estimated cross-year stability coefficients for the MET study, which collected four lessons per year per teacher, at 0.3 to 0.4 for most observation instruments. Clearly, cross-year stability will be affected by within-year scoring design, as error-filled within-year estimates will result in lower adjacent-year correlations.

Student surveys. Research on the stability of student surveys aggregated and used as teacher evaluation instruments is scarce. In the MET study, the stability of scales from the TRIPOD student survey instrument (Ferguson, 2008) from December to March in the same school year ranged from 0.7 to 0.85 (Kane & Cantrell, 2010), but this statistic was reported after correcting for measurement error. Polikoff (2013) used the MET data and found uncorrected, cross-year correlations in the 0.3 to 0.4 range.

Explanations for intertemporal stability. Some scholars have examined potential explanations for between-year differences in teacher quality metrics. Two separate studies using panel data from North Carolina (Goldhaber & Hansen, 2013; Jackson & Bruegmann, 2009) found teachers' value-added scores can be modestly predicted by the value-added scores of their peers. Goldhaber and Hansen further showed that peer absences predicted teachers' VA scores, and found small associations in the expected direction between teacher value-added scores and class size, student free lunch eligibility, percent of the class that is minority, and teacher absences. Papay and Kraft (2011), also using panel data from North Carolina, isolated an effect of teacher experience on changes in value-added scores.

In sum, the review of the literature suggests that more estimates of cross-year stability in teacher effectiveness metrics – particularly observational metrics and student surveys – would be useful, especially in light of the use of many of these metrics in recent teacher evaluation systems. As well, efforts to explain variability in cross-year metrics, and in particular value-added scores, may shape both the interpretation of these metrics as well as efforts to shape policies to improve teaching and learning. In this study, we add to this literature by examining adjacent-year correlations in key teacher accountability metrics, and explore whether differences in teacher scores across years can be explained by teacher perception, behavior, or classroom demographic information.

Data

We draw from data collected over three school years for the National Center for Teacher Effectiveness study. The study investigated the relationships between teacher characteristics, instruction, and achievement for a sample of fourth- and fifth- grade elementary math teachers and their students from four urban East Coast public school districts.

The sources of data in the study included: (1) up to three videos of instruction from each teacher per year, scored by two raters on the Mathematical Quality of Instruction (MQI) observation instrument (Hill et al., 2008), and by one rater on the Classroom Assessment Scoring System (CLASS) observation instrument[3] (Pianta, LaParo, & Hamre, 2007); (2) teacher surveys administered twice per year, with questions capturing teacher knowledge, beliefs, behaviors, and background; (3) TRIPOD (Ferguson, 2008) student surveys administered once per year in the spring, with questions regarding student background and student perceptions of mathematics classrooms and teachers; and (4) student administrative data, including standardized state test scores, student scores on an alternative mathematics assessment administered by the project, and demographic information.

In this paper, we focus our attention on cross-year stability of three metrics currently in use for teacher evaluation purposes: teacher value-added scores derived from student test data, scores derived from the application of classroom observation instruments, and aggregated student reports of classroom quality. These teacher quality measures were selected because of their widespread use in teacher evaluation systems (Herlihy et al., 2013), academic interest in such measures (Kane & Staiger, 2012), and demonstrated impacts on student outcomes (Brophy & Good, 1986; Chetty, Friedman, Hilger, Saez, Schanzenbach, & Yagan, 2011; Kane & Staiger, 2012).

To examine the predictors of teachers' scores across years, we selected variables either previously shown (Goldhaber & Hansen, 2013) or theorized to potentially explain variability in teacher quality metrics. These predictors include changes in classroom demographic composition, teacher self-reported coaching and professional development experiences, and changes in the school environment (increased test preparation; school resources) that might impact teacher quality. To assess correspondence between teachers' and objective indicators of classroom quality and student learning, we also measured and included teachers' perceptions that the academic quality and behavior of their students had declined or improved since the prior year.

[3] We chose to observe three lessons per year because of results of a prior decision study (Hill, Charalambous, & Kraft, 2012) and because three is likely similar to the number of observations enacted in many teacher evaluation systems.

Tables 1a and 1b describe the measures and predictors used in the study, and including the average internal-consistency reliabilities for variables composed of multiple items.

Table 1a. Teacher Quality Measures

Measures derived from videotaped observations

Mathematical Quality of Instruction (MQI) Measures (Hill et. al, 2008)

· *Richness* captures the sense-making and mathematical practices present in a teacher's instruction (6 items, $\bar{\alpha} = .59$)

· *Errors* captures the prevalence of teacher errors, imprecision, or a lack of clarity in a teacher's instruction (3 items, $\bar{\alpha} = .64$)

· *CCSP* captures the prevalence of Common Core mathematics-aligned student practices during a teacher's instruction (3 items, $\bar{\alpha} = .82$)

Classroom Assessment Scoring System (CLASS) Measures (Pianta et. al, 2007)

· *Emotional Support* captures the level of positive climate, sensitivity, and regard for student perspectives demonstrated in a teacher's instruction (3 items, $\bar{\alpha} = .79$)

· *Classroom Organization* captures the level of negative climate (reversed), classroom productivity and behavior management activities present in a teacher's instruction (3 items, $\bar{\alpha} = .72$)

· *Instructional Support* captures the quality of feedback and instructional dialogue, instructional learning formats, and the focus on students' content understanding, analysis, and inquiry in the teacher's instruction (5 items, $\bar{\alpha} = .87$)

Measures derived from TRIPOD

· *TRIPOD 7Cs* captures student perceptions of the math s/he is doing in the classroom, of the teacher's ability to teach math, and the environment created by the teacher for learning math (26 items, $\bar{\alpha} = .90$)

Value-Added indicators

· *State Test* captures the teacher's impact on student achievement on the state standardized math test

· *Alternate Test* captures the teacher's impact on student achievement on an alternative assessment more aligned with the Common Core Standards for math

Table 1b. Predictors of Teacher Quality Measures

Predictors

· *Coaching and collaboration* captures teachers' self-reported frequency of work with math coaches and other teachers the previous year (5 items, $\bar{\alpha} = .87$)

· *Professional Development* captures the teacher self-reported time spent on math-related learning the previous year (6 items, $\bar{\alpha} = .86$)

· *Change in Test Prep Behaviors* captures the change in teacher self-reported time spent on test preparation activities or instruction from the previous year to the current year (10 items, original variable $\bar{\alpha} = .77$)[4]

· *Change in School Resources* captures the change in teacher perceptions of the resources provided by his/her school (autonomy, enjoyment, teaching materials, professional development, freedom from interruptions in instruction) from the previous year to the current year (9 items, original variable $\bar{\alpha} = .66$)[5]

· *Change in FRPL* captures the percent difference of students eligible for free- or reduced-price lunches in a teacher's current year classroom as compared to the previous year

· *Change in LEP* captures the percent difference of Limited-English Proficiency students in a teacher's current year classroom as compared to the previous year

· *Change in SPED* captures the percent difference of special education students in a teacher's current year classroom as compared to the previous year

· *Change in Students' Base Achievement* captures the change in the average state math test achievement in a teacher's current year classroom compared to the previous year

· *Teacher perceptions of students' ability* captures teacher agreement with statements such as "In general, students in this year's class have more learning difficulties than students in last year's class" (reversed) and "This year's class has fewer behavior problems than last year's class." Higher scores indicate teachers perceive higher-ability and better-behaved students. (4 items, $\bar{\alpha} = .85$)

Video and student survey scores were created by averaging responses to items across lessons and students within a year, respectively.[6] To recover single-year teacher-level scores, we estimated the following multilevel model:

$$Y_{jky} = \beta_0 + \mu_{ky} + \epsilon_{jky,e}$$

where the outcome, Y_{jky}, represents the lesson-level j or student-level j MQI, CLASS, or TRIPOD 7Cs score in year y. Our model takes into account the nested structure of our data, with either lessons being nested within teachers, or students being nested within teachers. The parameter μ_{ky} represents teacher k's random effect on Y_{jky} in year y. Each teacher k's random

[4] The reliability estimate represents the internal consistency of the items generating the test prep behavior composite, not the reliability of the change.

[5] See note 2.

[6] MQI dimensionality was based on prior multilevel item response theory exploratory and confirmatory factor analyses (Kelcey, McGinn, Hill, & Charalambous, 2014). CLASS dimensionality was based on suggested structures by instrument designers. TRIPOD dimensionality was informed by exploratory factor analysis suggesting a single latent trait loading onto all items.

effect μ_{ky} represents his or her score on MQI, CLASS, or TRIPOD 7Cs in year y, adjusted for differences in reliability due to differences in number of lessons j or students j taught.

Value-added scores for each teacher were constructed using a multilevel model which controlled for student-level prior achievement and demographic indicators, but not peer- or cohort-level aggregates for these covariates.[7] Similar value-added models have been employed by vendors for the District of Columbia, Pittsburgh, and Florida (Goldhaber & Theobald, 2012).

Teacher-level scores for predictors of teacher a quality measures were simple averages of items within a year, or differenced values between years when appropriate (i.e. the "Change" predictors).

Sample

To conduct our stability analyses, we restricted the dataset to: (1) teachers who had at least two consecutive years of scores for all investigated measures; (2) teachers who had prior year measure scores for each year; and, (3) teachers who had data for each predictor of teacher quality for the given year[8]. Because our current dataset is incomplete with regard to student administrative data, two separate samples were created, one for instructional quality as rated from videos and reported by students (n=181, up to three years of data) and one for value-added scores (n=150 teachers, two years of data).

Analyses

To arrive at a metric for the stability of each measure of teacher quality, we first estimate the following model:

$$Y_{ky} = \chi_y + \mu_k + \epsilon_{ky,e}$$

where the outcome, Y_{ky}, represents teacher k's score in year y for the measure of interest. The parameter χ_y represents a vector of fixed effects for the year y in which the score was measured for the teacher; this fixed effects vector controls for differences between years in the average teacher score and distribution of scores for the measure. Controlling for this vector of fixed effects, the parameter μ_k represents the random teacher k effect on Y_{ky}, and the parameter $\epsilon_{ky,e}$ represents the residual of Y_{ky}. The variances of these two parameters are used in our estimation of cross-year stability for teacher measures, using the following equation:

$$\rho = \frac{var(\mu_k)}{var(\mu_k) + var(\epsilon_{ky,e})}$$

[7] For the full specification of the value-added models used, see the Appendix.

[8] A small number of teachers had incomplete sets of data for predictor variables in any given year. For these teachers (Year 2, n=4; Year 3, n=4), their scores for the missing predictor variable was mean-imputed. An imputed variable dummy was subsequently used in analyses.

The outcome, ρ, which represents measure stability, reflects the percentage of total variance in the average teacher k's score in year y on the measure that is attributable to differences between teachers on their effects on their observed scores. Importantly, this component of variance represents true score, or actual teacher ability, variance separate from two contributing factors to measure instability: cross-year changes to all teachers' performances on measures due to differences between years (i.e. the fixed effect vector χ_y, which might, for example, capture if all lessons are comparatively scored in a specific year more leniently due to changes in observational instruments or raters) and cross-year changes to teachers' performances due to differences between teachers idiosyncratic to specific years (i.e. $\epsilon_{ky,e}$, which might, for example, capture whether a specific teacher performed poorly in classroom observations one year due to a particularly misbehaving classroom).

We chose to estimate measure cross-year stability with these two equations instead of calculating cross-year correlations for two primary reasons. First, some teachers persisted in our observational and TRIPOD datasets over three years, and these multi-level models correct for the presence of two separate cross-year relationships for those teachers. Second, the variance components outputted from the multilevel framework of the estimated equations provide insight as to how much the average teacher's score in a given year can be attributable to actual differences between teachers in terms of ability as opposed to differences in ability due to idiosyncratic differences between teachers in specific years. Further analyses can then subsequently explore what percent of the variance due to the latter factor, the interaction of teacher and year, can be explained by individual predictors varying by teacher and year.

Differences in teacher performance on measures of teacher quality between years can also be a result of other factors specific to the teacher in a given year. To investigate the impact of each predictor on teacher measure scores from one year to the next, we estimate the following model:

$$Y_{kt} = \beta Y_{kt-1} + \delta V_c + \chi_y + \mu_k + \epsilon_{ky,e}$$

where the outcome of interest, Y_{kt}, represents teacher k's score at time t on the measure of interest. The parameter βY_{kt-1} represents the effect of teacher k's score on the outcome of interest Y_{kt} at time *t-1*[9]. This parameter captures the impact of a teacher's prior year ability on his or her performance on the measure in the current year. Similar to the multilevel equation used to estimate stability, the parameter χ_y represents a vector of fixed effects for the year y in which the score was measured for the teacher. Controlling for this vector of fixed effects, the parameter μ_k represents the random effect of teacher k on Y_{ky}, and the parameter $\epsilon_{ky,e}$ represents the residual of Y_{ky}.

δV represents a vector of predictor variables that might impact teacher performance on measures of quality and their regression coefficients, varying from model to model. Using variables from Table 1b, the different model vectors include: (1) teacher resource model, including the

[9] An alternative way of modeling this equation is to treat the dependent variable as a 'difference score', $Y_{kt} - Y_{kt-1}$. Conclusions from this equation would be interpreted as causes for the *change* in teacher scores from one year to the next, but it would also constrain the coefficient of teacher prior ability on teacher current ability at 1.

predictors *Coaching and Collaboration, Professional Development, Change in Test Prep Behaviors, and Change in School Resources*; (2) student demographic model, including the predictors *Change in FRPL, Change in LEP, Change in SPED, and Change in Students' Base Achievement*, and; (3) *Teacher Perceptions of Student Ability*.

From the regression coefficients of the vector of predictors in each model estimated, we can arrive at the estimated impact of each predictor on each measure of teachers and teaching quality *after* controlling for prior year performance of the analyzed measure.

Results

Table 2 below displays the stability of each measure considered in our analysis.

Table 2. Cross-Year Stability of Teacher Quality Measures

Measure	Average Within-Year ICC	Cross-Year Stability (ρ)
Video Measures, Teachers=181		
CWCM	.09	.28
Richness	.18	.39
Errors	.15	.35
CCSP	.31	.35
Emotional Support	.15	.46
Classroom Organization	.19	.17
Instructional Support	.07	.05
Student Survey, Teachers=181		
TRIPOD 7Cs	.16	.52
Value-Added Measures, Teachers=150		
State Test	.19	.50
Alternate Test	.09	.25

From Table 2, we see that none of the measures of teacher quality demonstrate high levels of cross-year stability, and that the stability statistics range from low ($\rho = .05$) to moderate ($\rho = .52$). State value-added scores and TRIPOD scores were the most stable across years, with the state value-added score stability among the highest documented in the existing literature and the TRIPOD stability considerably higher than found during the MET study (Polikoff, 2013). Several observational dimensions also showed marked persistence across years, including CLASS' Emotional Support Dimension and MQI's Richness dimension. Other dimensions, most notably those that capture classroom productivity and student behavior (CWCM, Classroom Organization) showed lower continuity across years. Cross-year stability was largely related to within-year ICC, with a correlation of roughly 0.40, suggesting that cross-year stability is a function of within-year precision of measurement. One exception to this trend was the relationship of the two statistics of MQI's CCSP dimension, which demonstrated levels of cross-year stability comparable to other measures despite showing a markedly higher within-year reliability. This suggests that, though within-year measurement error may contribute to instability, other additional factors influenced the measure's cross-year measure stability. Interestingly, cross-year stability coefficients typically exceeded ICCs, suggesting that the latter metric may underestimate the amount of true-score variance in teacher scores.

Table 3 below shows the regression coefficients, after controlling for prior year performance, for each predictor on current year measures of teacher quality from our analyses.

Table 3. Regression Coefficients for Predictors of Teacher Quality Measures

	Model 1				Model 2				Model 3
Measure	Coaching and Collaboration	Professional Development	Change in Test Prep Behaviors	Change in School Resources	Change in FRPL	Change in LEP	Change in SPED	Change in Students' Base Achievement	Teacher Perceptions of Student Ability
Video Measures									
CWCM	.119~	-.070	-.023	-.008	-.215	-.110	.170	.215	.161**
Richness	.068	-.014	-.130*	.066	-1.031*	.577	-.735	-.001	.081
Errors	.046	-.112	.028	.098	.675	-.639	-.191	.637**	.001
CCSP	-.057	-.099	.024	.093~	-.918*	.309	-.445	.260	.076
Emotional Support	-.030	.051	.074	-.060	-.439	-.397	.663	.066	.056
Classroom Organization	-.05	.035	.079	-.082	-1.370*	.254	.852	.296	.122*
Instructional Support	-.042	.043	.016	-.066	-.063	-.511	.172	.126	.040
Student Survey									
TRIPOD 7Cs	.050	.031	-.015	.102~	-1.063*	.409	.086	.226	.216**
Value-Added Measures									
State Test	.088	.050	.104	.178*	-.919*	.335	-.610	-.057	.142*
Alternate Test	.140	-.034	.015	.111	-.547	-.011	.042	.280	.182*

*Note: ~p<.10 *p<.05 **p<.01*. All measures and predictors have been standardized *except* for the predictors in Model 2, which are scaled in percentages. For example, a one standard deviation increase in a teacher's test prep behaviors from one year to the next is associated with a .130 standard deviation decrement in a teacher's current Richness score, on average in the population, after controlling for the teacher's scores on the other predictors of the model, the teacher's prior Richness score and the year of measurement.

WORK IN PROGRESS: Please do not circulate or cite without permission.

Several trends emerge from Table 3. First, changes in the population of free- or reduced-price lunch (FRPL) eligible students in a teacher's classroom predicts a teacher's quality of instruction (i.e. Richness, CCSP, Classroom Organization), a teachers' TRIPOD 7Cs score, and teacher state test based value-added scores. For each of these cases, a 10-percent increase in FRPL-eligible students from the previous year to the current year was associated with approximately 0.1 standard deviation decrease in the analyzed measure, on average in the population. Neither changes in the special education population or English Language Learners appeared associated with any of the outcome variables.

Second, when teachers viewed the students they currently taught more positively as compared to the students they taught in the previous year, their instruction was more often rated as connected to mathematics (CWCM). Students' TRIPOD 7Cs reports, aggregated to the teacher level, were also higher in this condition, indicating that teachers' and students' assessments of one another converged. Better views of students as compared to the prior year also predicted academically meaningful increases in their value-added scores across years. The relationship between measures of teacher quality and teacher perceptions is interesting. In the case of the measures that at least partially capture student behavior, it suggests that teachers, raters, and students tend to agree regarding this dimension of classroom quality. In the case of the value-added measures, it suggests that teachers are prescient regarding their value-added scores that will result from their current students' testing. This both validates value-added scores – teachers are reporting trends similar to those seen in test score data – but also calls into question their use in high-stakes evaluations, as similar instruction across years (at least on some observation-based dimensions) yields different results. However, the directionality is of this relationship is unclear; teachers who see their students as worse than previous years may be less motivated to provide strong instruction, resulting in lower scores on the analyzed measures.

Finally, we generally failed to find significant associations between teacher reports of professional development activities, school resources, and test preparation activities and the teacher quality measures used in Model 1. For the few cases where a significant relationship was observed, the direction of the relationship matched expectations. Self-reported increases in test preparation behaviors resulted in instruction that was less mathematically rich, and self-reported increases in access to school resources resulted in better performance on state value-added measures.

Conclusion

In this paper, we find that the cross-year stability in teacher accountability metrics is consistent with that found elsewhere in the research literature. Value-added scores based on the state test and TRIPOD 7Cs scores, averaged from student reports of classroom quality, were the most stable. Classroom dimensions associated with teachers' presentation of content and classroom climate were relatively more stable than those that capture classroom behavior and productivity dimensions.

The stability of the teacher state value-added scores is consistent with that found in extant literature. Used extensively in teacher evaluation systems across the US in response to federal policies, teacher value-added scores and their stability have become an oft-debated topic in the realm of education. As a result, it is notable that our results suggest similar levels of stability for some measures of teacher instruction (i.e. Richness, CCSP), whose many sources of variance (i.e. number of raters, number of lessons) have often resulted in lower measure reliabilities (Kane & Staiger, 2012).

In exploring predictors of cross-year scores, we found that some of the instability in the classroom behavior dimensions may be associated with between-year changes in classroom composition. Teachers' perceptions of student ability negatively correlated with MQI's Classroom Work Connected to Mathematics, and student FRPL status negatively correlated with CLASS' Classroom Organization. Notably, FRPL status also negatively correlated with two dimensions that capture the disciplinary integrity of the mathematics (Richness and Common Core-Aligned Student Practices). The causal mechanisms within these relationships are hard to parse, as students themselves may be more or less inclined to participate productively in classroom mathematics as a result of their backgrounds and experiences; teachers themselves may also adjust instruction based on the background of students. However, our findings suggest that some component of the cross-year deviations in teacher scores reflect real changes in classroom conditions as opposed to measurement error

Interestingly, teachers' own perceptions of their students' ability does predict the classroom behavior dimensions (CWCM and Classroom Organization) as well as TRIPOD 7Cs and VAM scores, suggesting that teachers and third party observers agree about differences between classes and how those differences may affect student performance on state tests. That teachers can predict change in student performance on standardized tests suggests that some of the cross-year instability is due to differences in classroom composition and/or the changes in instruction that teachers make as a result.

Finally, few of teachers' reports of professional learning opportunities predicted changes in scores on the classroom observation or student-based metrics. School resources, which is composed of a battery of items asking about school-supplied resources, professional development, professional autonomy and satisfaction with the school environment, was the exception.

References

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

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. *Value added models in education: Theory and applications*, 272-297.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Brophy, J. E. (1973). Stability of teacher effectiveness. *American Educational Research Journal*, 245-252.

Brophy, J. E., Coulter, C. L., Crawford, W. J., Evertson, C. M., & King, C. E. (1975). Classroom observation scales: Stability across time and context and relationships with student learning gains. *Journal of Educational Psychology*,*67*(6), 873.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593-1660.

Ferguson, Ronald. F. 2008. The TRIPOD Project Framework. Cambridge, MA: Harvard University.

Goldhaber, D., & Hansen, M. (2013). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*, *80*(319), 589-612.

Goldhaber, D., & Theobald, R. (2012). Do different value-added models tell us the same things. *Carnegie Knowledge Network*.

Good, T. L., & Grouws, D. A. (1975). Process-Product Relationships in Fourth Grade Mathematics Classrooms.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2013). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85-108.

Kane, T., & Cantrell, S. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *MET Project Research Paper, Bill & Melinda Gates Foundation*, 9.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kelcey, B., McGinn, D., Hill, H. C., & Charalambous, C. Y. (2014). Dimensionality and Generalizability of the Mathematical Quality of Instruction Instrument. Paper to be presented at the 2014 Annual Meeting of the National Council on Measurement in Education, Philadelphia, P.A.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. National Center on Performance Incentives, Vanderbilt, Peabody College.

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

Marshall, H. H., Green, J. L., Hartsough, C. S., & Lawrence, M. T. (1977). Stability of classroom variables as measured by a broad range observational system. *The Journal of Educational Research*.

Papay, J. P., & Kraft, M. A. (2011). *Productivity returns to experience in the teacher labor market: methodological challenges and new evidence on long-term career growth*. Working Paper.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2007). Classroom Assessment Scoring System (CLASS) Manual. Baltimore, MD: Brookes Publishing.

Polikoff, M. S. (2013). The stability of observational and student survey measures of teaching effectiveness. Paper presented at the 2013 Annual Conference of the Association for Education Finance and Policy, New Orleans, LA.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Appendix

Value-Added Model

To calculate value-added scores for each teacher in any given year, we estimate the following equation:

$$a_{jckgdt} = \beta_0 + \beta_1 A_{jt-1} + \beta_2 X_{jt} + \gamma_{gt} + \delta_d + \mu_k + \epsilon_{jckgdt}$$

Where the outcome of interest, a_{jckgdt}, represents student j's standardized score on either the state or alternate mathematics exam at time *t*;

A_{jt-1} represents a vector of prior achievement for student j in time *t*-1, including a linear, quadratic, and cubic term for student j's mathematics exam score at time *t*-1, and a linear term for student j's score on the reading exam from time *t*-1;

X_{jt} represents a vector of student demographic indicators for student j at time t, including gender, race, free- or reduced-price lunch eligibility, special education status, and limited English proficiency; and

Also included in the model are district fixed-effects, δ_d, and a vector of grade-by-year fixed effects, G_{gt}, to account for differences across grades and school years.

To be included in the model, student j's tested grade at time *t* must follow sequence with regards to his or her tested grade at time *t*-1. Furthermore, student j's class *c* must have fewer than 50% of students having special education status, fewer than 50% of students missing scores for the prior achievement vector, and, after all other restrictions, must have a sample of at least five students.

In our model, students are nested within teachers; thus, we include a random effect μ_k in the multilevel model. The estimated teacher effect $\widehat{\mu_k}$ represents teacher k's value-added score, the empirical Bayes estimate of the random effect that is a best linear unbiased prediction. These estimates are "shrunken" estimates, which account for differences in the reliability of the estimates from teacher to teacher by shrinking less reliable estimates toward the mean. This shrinkage reduces random error that is associated with the class- and student-levels, including error due to small samples of students.

Many debates in the literature have revolved around the bevy of possible modeling options for value-added scores. We consciously chose to exclude from our model peer and cohort effects (i.e. aggregation of student demographics and achievement variables at the class and school level), as one of our predictors of teacher quality measures involved changes in these covariates from year to year.