Comparing the Alignment between Two Observational Measures of Science Teachers' Instructional Practice

Jamie N. Mikeska, Joshua M. Rosenberg, Steven Holtzman, S., and Daniel McCaffrey

Problem

Teacher evaluation systems are currently in use within all 50 states across the United States (Kane, Kerr, & Pianta, 2014). These evaluation systems are designed to address two main purposes: (1) to enable school district leaders to make promotion, retention, and hiring decisions for K-12 teachers and (2) to provide formative feedback to these teachers in order to improve their instructional practice (Darling-Hammond & Hamilton, 2012). Although the individual components within these systems and how they are used to support these two purposes varies within districts, one of the consistencies across all these systems is the common use of observational measures to assess teaching quality and provide formative feedback to teachers (National Council on Teacher Quality, 2012). For example, Danielson's Framework for Teaching observation protocol, a generic instrument that assesses teachers' practice on four teaching domains (e.g., classroom environment, instruction), is currently being used in many school districts nationwide as part of their formal teacher evaluation systems (McGuinn, 2012).

Due to the widespread use of observational measures within these evaluation systems, there has been a significant increase in research studies examining the reliability and validity properties of these types of teaching quality measures. A majority of this research has focused on examining the influence of rater characteristics and sampling strategies on the accuracy and reliability of teachers' observation scores (Measures of Effective Teaching (MET), 2013). For example, researchers have found that using more than one rater and more than one lesson increases reliability (MET, 2013), administrators are better able to differentiate among teachers than peer raters (MET, 2013), coding an amount of time greater than 30 minutes does not significantly increase the reliability of observations (Joe, McClellan, & Holtzman 2014), and modality of lesson sampling (e.g., live or pre-recorded videos) has little impact on observational scores (Casabianca, McCaffrey, Gitomer, Bell, Hamre, & Pianta, 2013). More recently some researchers have argued for the importance of attending to the subject-specific nature of teachers' instructional practice (Charalambous, Komitis, Papacharalambous, & Stefanou, 2014; Hill & Grossman, 2013). Yet, research has not taken into account how differences in the actual observation protocols used may impact the conclusions drawn about individual teachers' instructional quality and the formative feedback provided to teachers about their strengths and areas for growth. Instead, these research studies and formal evaluation systems tend to select one observational measure for use. Ideally, one would hope that the observational protocols used would provide consistent signal about which teachers are more or less effective in the classroom. However, it is unclear to what extent this is the case.

To address this gap, this study examines the alignment between two science-specific observational measures of teaching quality -- the Quality Science Teaching – Measures for Effective Teaching (QST-MET) protocol and the Inquiring into Science Instruction Observation Protocol (ISIOP) – in the context of secondary biology teachers' instruction. In particular, we investigate the extent to which these two observational instruments would categorize science

teachers similarly within an evaluation system. Understanding to what extent various observational measures classify teachers in similar ways and provide consistent feedback on their teaching practice can help districts and administrators make better choices about which observational instruments to use (Marshall, Lotter, Smart, & Sirbu, 2010). Most importantly, it is imperative to ensure that teachers' rankings within these evaluation systems are not subject to undue volatility from the observational protocol used to assess teachers' practice.

Research Methods

Sample

The sample for this study consisted of 99 ninth grade biology teachers who participated across two years of the MET study. Participating teachers (72% Caucasian and 67% female) work in five school districts and reported a range of teaching experience (31% have taught for three years or fewer, while 29% have more than 10 years of classroom experience). Overall, the background characteristics of the 99 biology teachers participating in this study, as well as those of their students, are similar to the full population of MET biology teachers and students. We randomly selected three to five of the video-recorded lessons per teacher, yielding 474 videos of classroom practice; 403 of these 474 lessons were previously scored with the QST-MET measure, which was developed by researchers at Stanford, as part of the MET study (Schultz & Pecheone, 2013).

Data Collection

Primary data collection involved 12 raters scoring 474 videos of the secondary biology teachers' lessons using the ISIOP. The ISIOP (Minner & DeLisi, 2010) was designed to provide a compilation of quantitative scores representing different features of a science teacher's instructional practice. For this study, we focus on teachers' instructional leadership practices (ILP; four domains – teaching style, support for self-directed learning, lesson organization, and dealing with distractions) and teachers' use of investigation-related experiences (IRE; two domains – student-directed IRE and teacher-directed IRE). The scores for these six domains were based on a four-point scale (0-3), with higher scores indicating greater emphasis of these practices in a teacher's instruction. Each domain included a subset of items rated on this four-point scale, and we averaged items within each domain to derive a score for each lesson on each of these six domains. We used data previously collected as part of the MET project using the QST-MET observation protocol. The QST-MET measures sought to assess teachers' instructional practice across three clusters, with four items per cluster, as shown in Table 1. Like the ISIOP, these QST-MET items were rated on a four point scale and higher scores indicated greater emphasis and/or higher quality of these practices. In summary, each lesson received 12 different scores using the QST-MET protocol and six different scores using the ISIOP.

Table 1. QST-MET Clusters and Items

Cluster 1	Cluster 2	Cluster 3

1a. Sets the context and focuses learning on key science concepts 1b. Uses representations 1c. Demonstrates content knowledge 1d. Provides feedback to students	2a. Promotes students' interests in and motivation to learn science 2b. Assigns tasks to promote learning and addresses the demands of the task for all students 2c. Uses modes of teaching science concepts 2d. Elicits evidence of students' knowledge and understanding	3a. Initiates the inquiry 3b. Provides guidelines for conducting the experiment and gathering data 3c. Provides guideline for conducting the experiment and gathering data 3d. Guides analysis and interpretation of data

Data Analysis

There were three main parts to the analysis: (1) a content analysis of the alignment between these observational measures, (2) a quantitative examination of the relationships between the two measures, and (3) an examination of how these teachers would be categorized within an evaluation system based on their scores on these observational measures. For the content analysis, we examined the descriptions of the six ISIOP domain scores and the 12 QST-MET item scores to determine which measures likely demonstrated a high degree of conceptual alignment. In the next step, we examined the bivariate correlations between the six ISIOP domain scores and 12 QST-MET item scores at the video level. Because the MET researchers only coded some of the videos for each QST-MET cluster, this part of the analysis included 270 videos for cluster 1, 280 videos for cluster 2, and 108 videos for cluster 3. Finally, to examine the practical significance of any differences noted between these scores, we examined changes in how teachers would be ranked from ineffective to highly effective using the two protocols. To do so, we grouped teachers into quartiles (quartile 1 – ineffective; quartile 2 –minimally effective; quartile 3 – effective; quartile 4 – highly effective) for each variable and then compared teachers' ranks across the different variables. Since there were a limited number of videos scored using the QST-MET cluster 3 measures, this final analysis focused on the relationship between teachers' rankings on the items in the QST-MET clusters 1 and 2 with the ISIOP ILP and IRE domains.

Findings

Due to proposal space limitations, we focus here on findings from the quantitative examination of the relationships between the QST-MET cluster 2 items (four scores) and the six ISIOP domains (4 ILP domain scores, 2 IRE domain scores). In addition, we highlight the practical significance of these findings within a teacher evaluation system. The full presentation and paper will include findings examining the full set of relationships across variables.

Relationships Between QST-MET and ISIOP Measures

Table 2 presents the correlations between QST-MET cluster 2 item scores and the 6 ISIOP domain scores. In general, this analysis reveals relatively weak relationships across these observational measures. The most notable finding was lessons that received higher ratings on the ILP teaching style were more likely to receive higher ratings on their ability to promote students'

interest and elicit students' understanding. However, overall even these associations were modest. In general, these findings suggests that these measures may be examining different features of teachers' instructional practice.

Table 2. Correlations Between ISIOP Instructional Leadership Practices (ILP) and Investigation-Related Experiences (IRE) Domain Scores and QST-MET Cluster 2 Items

	QST 2a: Promotes students' interests in and motivation to learn science	QST 2b: Assigns tasks to promote learning and addresses the demands of the task for all students	QST 2c: Uses modes of teaching science concepts	QST 2d: Elicits evidence of students' knowledge and understanding
ISOIP: Teaching Style	.296*	.150*	.173*	.320*
ISIOP: Support for Self-Directed Learning	-.016	.143*	.163*	-.051
ISIOP: Lesson Organization	.064	.068	.109	.044
ISIOP: Dealing with Distractions	.074	.035	.031	.034
ISIOP: Student-directed Activities	.097	.122*	.190*	.101
ISIOP: Teacher-directed Activities	.041	-.009	.024	-.003

* $p < .05$

Comparing Teachers' Rankings in an Evaluation System

Next, we present findings from our analysis about the practical significance of the use of different measures in terms of how teachers would be ranked in an evaluation system. To do so, we created quartile rankings for teachers on each of the variable scores. Then we compared these rankings across variables to determine whether their rankings would increase, decrease, or remain unchanged, depending on which measure was used to evaluate their teaching practice.

Overall, we found that the majority of teachers would be ranked differently using the observational measures across the two protocols. In particular, within each comparison, we found that approximately 70% of the teachers' rankings would change if the instrument used changed. To illustrate these findings in more detail, in Figure 1 we show the results of comparing how teachers' rankings would change for the relationship that demonstrated one of the strongest associations – the QST-MET cluster 2a item "Promotes student' interests in and motivation to learn science" and the ISIOP Teaching Style domain (blue bars) – and one that demonstrated a weak association – the same QST cluster 2a item and the ISIOP Teacher-directed Activities (TDA) domain (red bars). For the first comparison (QST-MET cluster 2a score vs. ISIOP Teaching Style score), this figure shows how approximately 40% of teachers in this sample would be rated as more effective if using the measure from the QST-MET. For the second

comparison (QST-MET cluster 2a score vs. ISIOP TDA score), this figure shows that about 1/3 of teachers would be ranked as more effective if using the QST-MET cluster 2a item, about 1/3 of teachers would be ranked as more effective using the TDA domain score, and about 1/3 of teachers would be ranked equally effective when using these two measures.

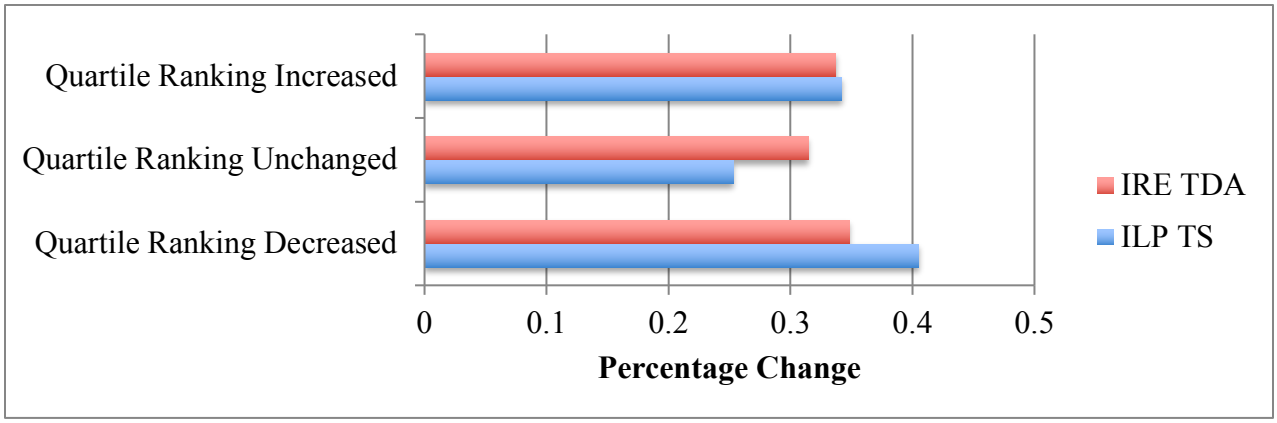

Figure 1. Percentage change in teachers' quartile rankings: Comparison with scores on QST-MET cluster 2a measure.

Contributions

In this study, we sought to explore the alignment between two science-specific observational measures of teaching quality, the QST-MET and the ISIOP. We found that the relationships between the two measures, using bivariate correlations, were small to modest. Only a few correlations were in the range of 0.20 to 0.40; overall, the majority of correlations were below 0.20. In terms of practical significance, we found that the majority of teachers' rankings would change across instruments; that is, it would be rare for teachers to be ranked similarly within an evaluation system using these two science-specific observation instruments.

Overall, these findings suggest that different observational measures, even those designed for similar purposes and disciplines, may not align and may measure different aspects of teachers' instructional practice. In terms of providing feedback for teachers, the use of different measures, then, may accentuate different aspects of teachers' instructional practice. We argue that administrators and other stakeholders may benefit from careful consideration of the focus of the observational measures they choose because the measure used is likely to impact (either positively or negatively) the conclusions drawn from their use. Moreover, measures should be carefully chosen in terms of the goals of the evaluation system as well as in terms of what other measures are used as part of the overall system. This work promises to be of interest to NARST members interested in measuring science teachers' instructional practice and those interested in the policy implications for using different observation measures within teacher evaluation systems.

References

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757-783.

Charalambous, C. Y., Komitis, A., Papacharalambous, M., & Stefanou, A. (2014). Using generic and content-specific teaching practices in teacher evaluation: An exploratory study of teachers' perceptions. *Teaching and Teacher Education, 41*, 22-33.

Darling-Hammond, L., Jaquith, A., & Hamilton, M. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching.* Stanford, California: Stanford Center for Opportunity Policy in Education (SCOPE). Retrieved October, 26, 2012.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371-384.

Joe, J.N., McClellan, C.A., & Holtzman, S.L. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In T.J. Kane, K.A. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (p. 415-443). San Francisco, CA: Jossey-Bass.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. San Francisco, CA: Jossey-Bass.

Marshall, J. C., Smart, J., Lotter, C., & Sirbu, C. (2011). Comparative analysis of two inquiry observational protocols: Striving to better understand the quality of teacher-facilitated inquiry-based instruction. *School Science and Mathematics*, *111*(6), 306-315.

McGuinn, P. (2012). *The state of teacher evaluation reform.* Center for American Progress.

MET Project. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's first three-years*. Seattle, WA: Bill & Melinda Gates Foundation.

Minner, D., & DeLisi, J. (2010). ISIOP technical report: Conceptual framework. Waltham, MA: Education Development Center, Inc.

National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies.* Washington D.C.

Schultz, S. E. & Pecheone, R.L. (2013). Assessing quality teaching in science. In J. Kane, K.A. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 444-492). San Francisco, CA, Jossey-Bass.