SREE Spring 2014 Conference Abstract Template

Abstract Title Page

Not included in page count.

Title: Observer Use of Standardized Observation Protocols in Consequential Observation Systems

Authors and Affiliations: Courtney A. Bell, Educational Testing Service; Qi, Yi, Educational Testing Service; Nathan D. Jones, Boston University; Jennifer M. Lewis, Wayne State University; Monica McLeod, Wayne State University; Shuangshuang Liu, Educational Testing Service

Background / Context:

Evidence from a handful of large-scale studies suggests that although observers can be trained to score reliably using observation protocols, there are concerns related to initial training and calibration activities designed to keep observers scoring accurately over time (e.g., Bell, et al, 2012; BMGF, 2012). Studies offer little insight into how educational practitioners understand and score observation protocols. This lack of clarity on the factors that facilitate and constrain educators' learning and use of observation systems makes it difficult to implement training and quality control processes at scale.

With a few exceptions (e.g., Cash, et al, 2012), existing research on the current generation of K-12 observations has conceptualized the work of observation primarily from a measurement perspective (e.g., BMGF, 2012; Grossman et al, 2010). By this, we mean the studies have conceptualized raters as observers who receive training that disciplines their view of teaching to match the observation protocol. Observers are then evaluated on the degree to which their scores are accurate and consistent with master observers' scores on the same lessons. Critical aspects of the scoring task have not been carefully investigated – for example, how the observer understands the task, the observers' ability to apply scoring criteria across lessons, subjects and grades, the observers' own beliefs about what counts as high quality teaching, or the relationship between the teacher and the observer.

Purpose / Objective / Research Question / Focus of Study:

In order to effectively train administrators at scale, it is critical to understand how administrators learn to complete two major tasks – learning to create accurate scores and learning to have conversations around those scores that support instructional improvement. Previous research has focused on the first of these tasks, but we argue that for principals, the two tasks are inextricably linked. This study takes the perspective that scoring observations of classroom interactions is a complex socio-cognitive process that must be understood in order to improve observer training, and ultimately, score quality. In the proposed session, we will present findings from the *Understanding Consequential Assessment Systems for Teachers (UCAST)* study, which investigates how principals, assistant principals, and other district personnel in Los Angeles Unified School District (LAUSD) learn to use a standardized observation protocol. Drawing on certification data from nearly 1000 administrators and think-aloud data from a subsample of 42 focus observers, this mixed-methods study describes which aspects of teaching were easiest and hardest for LAUSD observers to learn and investigates how observers used the observation protocol to score lessons.

We focus on two research questions: 1) What areas of the observation protocol were challenging for observers to learn to score accurately? And 2) Once trained, how did observers use the observation process?

Setting and Participants:

LAUSD, the site of the study, is the second largest public school district in the country, with more than 800 schools and a student enrollment of approximately 670,000. The student population is racially and ethnically diverse; teachers are similarly diverse. More than 76 percent of students are eligible for free/reduced lunches.

In the 2012-2013 school year, LAUSD introduced its new teacher evaluation system, known as the Teacher Growth and Development Cycle (TGDC). As part of the initial implementation of the TGDC, 998 administrators were trained to use the district's observation

system. Of the observers trained in the 2012-2013 school year for whom we have survey data, the vast majority (88% altogether) were principals (64%) or assistant principals (24%). Thirty-one observers were central office staff, which is about 5% of all the observers trained. Central office staff includes individuals who have job titles such as specialized director or Least Restrictive Environment (LRE) specialist. A small proportion of observers were instructional directors (1%). On average, the administrators who participated in training had 6 years of experience. The majority of administrators in the sample had their teaching certificate, and they were certified in a wide range of areas (Table 1), most in elementary education.

The 42 observers in this study are similar to the overall sample of 998 principals. Over half of the focus observers were principals, with another quarter serving as assistant principals, and 15% serving in other roles (e.g., coaches, directors, etc.). They had an average of 4.4 years in their current professional roles, including time in LAUSD or any other district. Among the focus observers, 38% work in elementary schools, 48% work in secondary schools, and 12% are instructional directors or coaches, who work with more than one school. Of the 32 focus observers for whom certification data were available, half were certified in elementary education; 34% of the focus observers were certified in English Language Arts, 25% held certificates for social studies, 25% held science credentials and 19% were certified in mathematics.

Intervention / Program / Practice:

The observation system on which administrators were trained was the Teaching and Learning Framework (TLF), a modified version of Danielson's *Framework for Teaching* (Danielson & McGreal, 2000). Danielson's original instrument is said to be the most widely used observation protocol in the country. In the summer before the 2012-2013 school year and during the school year, LAUSD provided a four-day training to support administrators. At the conclusion of training, observers needed to pass a certification exercise.

To become certified, observers watched a videotaped classroom lesson and collected evidence that was objective, detailed, and appropriately used to support scores. Observers needed also to be able to accurately score teaching practice, which is measured by their agreement with master observers' scores. They must be able to do this for all of 21 elements used in 2012-2013. Certification status is broken down into four categories of proficiency. If a principal scored in the lowest category, they were not allowed to perform observations.

Research Design:

This study draws on certification data from 998 administrators. Those data include the master scores and the administrative scores created as a part of the certification test administered at the end of training. We also draw on data from think-alouds and interviews of 42 focus observers. To better understand how observers use the TLF, they were asked to score a 10 minute video of a teacher. There were asked to work as they normally would, thinking out loud where possible. All observers thought aloud during the scoring parts of their work. After the observer completed all scoring, a stimulated recall session was conducted in which researchers asked specific questions about how the observer was thinking about specific scales (e.g., how she decided on a particular score, or why a certain score could not be higher or lower than what the observer assigned). All sessions were audio recorded.

Data Collection and Analysis:

Certification data were collected during the 2012-2013 school year. Think-aloud and interview data were collected in the spring of 2013. Mixed methods were used to analyze the various data sources. Descriptive statistics were used to analyze the certification data. A constant comparative method (Strauss & Corbin, 1998) was used to develop a grounded theory for analyzing the think-aloud and interview data.

Findings / Results:

As previously described, we conducted our analyses with two goals in mind—understanding where administrators struggled in the certification process and how they might use the observation process to improve their teachers' instruction. Below, we describe our findings of the accuracy of observers' scores, and then we describe how observers thought about and used the observation protocol.

Certification Data

Observer accuracy was assessed by comparing administrators' scores on the observation protocol to scores assigned by master raters. In general, all agreement rates were lower than desirable. The highest levels of agreement were in Standard 2, which is principally concerned with behavior management, organization, and the classroom environment (Table 2). The lowest levels of agreement – across all groups – were with elements in Standard 1 (Planning) and 5 (Professional Growth). On the standards concerning planning, standards-based instruction, and professional growth, observers tended to score teachers higher than master observers. On the remaining standard (Designing Instruction), they tended to score teachers slightly lower than master observers. Finally, principals had higher agreement rates than assistant principals; elementary and secondary administrators had similar agreement rates (Table 3 and 4).

Think-Aloud Data

Principals have many tasks related to observing. One is to create the scores. But another is to have conversations about those scores with teachers for the improvement of practice. One might reasonably think that these are two separate tasks – an observer creates a score and then has a conversation about the score. The think-aloud data suggest observers carry out these tasks in a more integrated way. Observers' scoring and reasoning processes suggest they use the TGDC with both evaluation and improvement in mind. This stands in contrast to research studies that have used researchers, not principals, to create scores (e.g., Bell, Qi, et al., 2013).

We found that when interviewing focus observers and watching them use the TLF in the think-alouds, they were often thinking about observing with the purpose of evaluation in mind. They were not thinking only about creating scores. We found that focus observers thought regularly about how they were going to have the post observation feedback session or the ongoing conversations across the school year. They also thought about how they might help a teacher improve her practice while they were observing and scoring.

Specifically, when observers were asked to explain why they gave the score they did, they often made reference to the outcomes of the scores – e.g., how the conversation with the teacher might go, what the observer might say to the teacher, what the observer's general approach is to the scoring conversation. For example, when one of our focus observers, Sara (a pseudonym), was describing how she would discuss the scores she gave to the teacher she said, "If this had been my teacher, I would have had him looking at the transcript. And we'd go down everything that was said. I'd tag with the teacher. It takes a long time."

Observers were very aware and discussed the fact that evaluation was a human endeavor and they had to consider politics and personalities when conducting observations. One observer, Anthony, explained that in this experience he had to think of the "human drama" a particular score is going to create.

Some individuals developed scoring processes to help them have productive evaluation conversations with teachers. For example, Heather's process of taking evidence and coding took account of how she will share the evidence and scores with her teachers. She explained the process she uses and noted that her coding work is used to help teachers see what she has done. Another common strategy observers used was to provide more evidence than they thought was strictly necessary. They explained that by showing the teacher that they (the principal) had been paying attention and taking careful notes, this would decrease the likelihood the teacher would feel they was not being objective and fair in the evaluation.

In addition to anticipating the post-observation conference or evaluation context, observers sometimes noted what the teacher should have done differently. This was most common when the observer was asked to justify, or explain why she gave a certain score, and occurred occasionally when the observer was actually watching the video of practice. For example, in describing the questioning technique the teacher used, an observer might note how the teacher could have gotten more students involved or how the teacher could have rephrased the question so that it was more cognitively challenging. As one focus observer (Ella) reported:

> At least in the 10 minutes, he did not have any classroom management problems. He had a few systems in place that could have been better -- in terms of picking up materials. And I saw that he lacked in terms of seeing that everyone had the materials. He should have had an overhead where everyone could read or be able to see. There were just a few things lacking.

This practice – of noting what the teacher should have done in order to justify a score – was one we saw rarely in previous think-aloud work with observers who were researchers, rather than principals, instructional directors, etc. (Bell, Qi, et al., 2013). Though we can only speculate, it is possible that LAUSD observers think about what the teacher could have done differently than researchers, in part, because they are responsible for helping improve instruction, where researchers are not.

Conclusions:

This study suggests administrators have a great deal of knowledge they bring to bear on the observation process. They also bring a commitment to improving instruction to their observation work. In other words, they are not blank slates as they go through observer training. That said, certification data suggest observers have much to learn about how to accurately score lessons according to the protocol. It is unclear whether the knowledge and commitments observers bring is supportive of high quality scores or the improvement of instruction. Future studies should investigate whether the way in which principals use observation protocols results in better, more useful observation scores.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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. doi: 10.1080/10627197.2012.715014

Bell, C.A., Qi, Y., Croft, A.C., Leusner, D., Gitomer, D.H., McCaffrey, D.F., Pianta, R. (2013). *Improving observational score quality: Challenges in observer thinking*. Unpublished manuscript.

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains*. Seattle, WA: Author.

Cash, A. H., Hamre, B.K., Pianta, R.C., & Myers, S.S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542. doi: 10.1016/j.ecresq.2011.12.006

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice* Alexandria, VA: Association for Supervision and Curriculum Development.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K. M., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores* (NBER Working Paper). Cambridge, MA: The National Bureau of Economic Research.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research*. Thousand Oaks, CA: SAGE Publications.

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

Table 1
Certification Areas of Observers Who Returned the Pre-Training Survey (N=677)

Certification area	Percentage	Count
Elementary education	63.4%	429
Math	13.7%	93
English/Language Arts	20.7%	140
Science	10.3%	70
Social Studies	17.0%	115

Note. Percentages do not add to 100% as some observers hold multiple certifications.

Table 2

Element-Level Reliability Statistics (Full Sample)

		% Exact Match (True Score)	Mean Deviation	% Exact Match (Other Observers)
Standard 1	*Planning and Preparation*	*42%*	*0.24*	*42%*
Element 1d1	Analysis & Use of Assess. Data for Planning	61%	-0.32	44%
Element 1d3	Standards-Based Learning Activities	52%	0.01	39%
Element 1d4	Purposeful Instructional Groups	65%	-0.25	47%
Element 1e1	Lesson and Unit Structure	30%	0.53	44%
Element 1e2	Aligns with Instructional Outcomes	41%	0.36	39%
Element 1e3	Criteria and Standards	30%	0.60	38%
Element 1e4	Design of Formative Assessments	17%	0.76	44%
Standard 2	*Designing Coherent Instruction*	*69%*	*-0.11*	*52%*
Element 2a1	Teacher Interactions with Students	69%	-0.02	53%
Element 2a3	Classroom Climate	66%	-0.02	49%
Element 2b2	Expectations for Learning and Achievement	69%	-0.34	51%
Element 2c1	Management of Routines, Procedures, and Transitions	70%	0.03	55%
Element 2d2	Monitoring and Responding to Student Behavior	72%	-0.19	55%
Standard 3	*Standards-Based Learning Activities*	*48%*	*0.12*	*44%*
Element 3a1	Communicating the Purpose of the Lesson	20%	0.70	48%
Element 3b1	Quality and Purpose of Questions	38%	0.49	39%
Element 3b2	Discussion Techniques	65%	-0.18	47%
Element 3c1	Standards-Based Projects, Activities, and Assignments	63%	-0.36	46%
Element 3c2	Purposeful and Productive Grouping of Students	60%	-0.20	42%
Element 3d1	Assessment Criteria	10%	1.08	38%
Element 3d3	Feedback to Students	67%	-0.21	49%
Element 3e1	Responds and Adjusts to Meet Student Needs	59%	-0.36	41%
Standard 5	*Professional Growth*	*43%*	*.23*	*32%*
Element 5a2	Use of Reflection to Inform Future Instruction	43%	0.23	32%

Table 3

Element-Level Reliability Statistics (By Job Role)

		Asst. Principal % Exact Match	Principal % Exact Match
Standard 1	*Planning and Preparation*	*41%*	*45%*
Element 1d1	Analysis & Use of Assess. Data for Planning	67%	60%
Element 1d3	Standards-Based Learning Activities	53%	56%
Element 1d4	Purposeful Instructional Groups	68%	66%
Element 1e1	Lesson and Unit Structure	23%	35%
Element 1e2	Aligns with Instructional Outcomes	36%	45%
Element 1e3	Criteria and Standards	28%	32%
Element 1e4	Design of Formative Assessments	13%	20%
Standard 2	*Designing Coherent Instruction*	*67%*	*72%*
Element 2a1	Teacher Interactions with Students	66%	73%
Element 2a3	Classroom Climate	62%	70%
Element 2b2	Expectations for Learning and Achievement	73%	70%
Element 2c1	Management of Routines, Procedures, and Transitions	64%	75%
Element 2d2	Monitoring and Responding to Student Behavior	68%	75%
Standard 3	*Standards-Based Learning Activities*	*47%*	*50%*
Element 3a1	Communicating the Purpose of the Lesson	15%	22%
Element 3b1	Quality and Purpose of Questions	32%	43%
Element 3b2	Discussion Techniques	69%	67%
Element 3c1	Standards-Based Projects, Activities, and Assignments	63%	67%
Element 3c2	Purposeful and Productive Grouping of Students	65%	61%
Element 3d1	Assessment Criteria	8%	11%
Element 3d3	Feedback to Students	62%	70%
Element 3e1	Responds and Adjusts to Meet Student Needs	60%	60%
Standard 5	*Professional Growth*	*37%*	*47%*
Element 5a2	Use of Reflection to Inform Future Instruction	37%	47%

Table 4

Element-Level Reliability Statistics (By Instructional Level)

		Elem. % Exact Match	Sec. % Exact Match
Standard 1	*Planning and Preparation*	*45%*	*41%*
Element 1d1	Analysis & Use of Assess. Data for Planning	64%	59%
Element 1d3	Standards-Based Learning Activities	55%	53%
Element 1d4	Purposeful Instructional Groups	68%	64%
Element 1e1	Lesson and Unit Structure	34%	25%
Element 1e2	Aligns with Instructional Outcomes	45%	40%
Element 1e3	Criteria and Standards	32%	30%
Element 1e4	Design of Formative Assessments	18%	16%
Standard 2	*Designing Coherent Instruction*	*71%*	*70%*
Element 2a1	Teacher Interactions with Students	71%	70%
Element 2a3	Classroom Climate	70%	64%
Element 2b2	Expectations for Learning and Achievement	71%	73%
Element 2c1	Management of Routines, Procedures, and Transitions	72%	72%
Element 2d2	Monitoring and Responding to Student Behavior	74%	72%
Standard 3	*Standards-Based Learning Activities*	*50%*	*48%*
Element 3a1	Communicating the Purpose of the Lesson	22%	15%
Element 3b1	Quality and Purpose of Questions	41%	38%
Element 3b2	Discussion Techniques	68%	67%
Element 3c1	Standards-Based Projects, Activities, and Assignments	65%	66%
Element 3c2	Purposeful and Productive Grouping of Students	60%	66%
Element 3d1	Assessment Criteria	11%	8%
Element 3d3	Feedback to Students	70%	66%
Element 3e1	Responds and Adjusts to Meet Student Needs	61%	60%
Standard 5	*Professional Growth*	*44%*	*46%*
Element 5a2	Use of Reflection to Inform Future Instruction	44%	46%