Abstract Title Page
Not included in page count.

Title: Predicting Observer Training Satisfaction and Certification

Authors and Affiliations: Courtney A. Bell, Educational Testing Service, Nathan D. Jones, Educational Testing Service, Jennifer M. Lewis, Wayne State University, Shuangshuang Liu, Educational Testing Service

Background / Context:

The last decade produced numerous studies that show that students learn more from high-quality teachers than they do from lower quality teachers (Aaronson, Barrow & Sander, 2007; Clotfelter, Ladd & Vigdor, 2007; Rivkin, Hanushek & Kain, 2005; Rockoff, 2004). In fact, high-quality teachers trump curriculum, institutional organization, class size, and any number of malleable school factors that affect the quality of instruction. A number of government and private sector initiatives are aiming to improve teacher evaluation systems as a lever for ensuring high-quality teaching (e.g., Bill & Melinda Gates Foundation (BMGF), 2010; 2011; U.S. Department of Education, 2010). These emerging models of teaching evaluation have focused primarily on two measures of teaching quality: observations and some measure of student achievement, most often value-added models (VAM) or student growth percentile scores.

If instruction is to improve through the use of more rigorous teacher evaluation systems, the implementation of these systems must provide consistent and interpretable information about which aspects of teaching practice need improvement and how those improvements can be accomplished. While there has been a great deal of research and commentary on the quality of VAM and other measures of student growth (e.g., National Research Council & National Academy of Education, 2010), there has been relatively little scrutiny given to the use of observation protocols in the context of evaluation.

A primary concern for using observation systems in teacher evaluation is the challenge of training observers to score in valid and reliable ways (Bell et al., 2012; BMGF, 2011; Casabianca, McCaffrey, Gitomer, Bell, & Hamre, 2012). Recent studies suggest that sizable proportions of observers struggle to be certified and they subsequently exhibit unacceptable levels of reliability and accuracy. Despite the potential challenges in training local personnel to serve as observers, districts and states are moving forward. Many districts understand the challenges they face and are concerned, yet existing research offers only modest insights and advice (e.g., Gitomer et al., 2012).

Purpose / Objective / Research Question / Focus of Study:

In the proposed session, we will present first-year findings from the *Understanding Consequential Assessment Systems for Teachers* (UCAST) study, which investigates how administrators in a large urban school district learn to use a standardized observation protocol. UCAST collects extensive data on more than 700 principals, assistant principals, and other district personnel being trained as observers. The study focuses on how observer background characteristics, understandings of the observation protocol, and training quality shape score reliability and validity. During this first year of data collection, we focus on the predictors of observer certification and training satisfaction, both of which contribute to the study's larger research agenda. The relevant research questions are as follows:

1. To what extent do administrator characteristics, beliefs, and expectations predict training satisfaction?
2. To what extent do administrator characteristics, beliefs, and expectations predict certification success?
3. What components of the observation protocol are most challenging for observers to certify, and what accounts for these challenges?

Setting:

This project takes place in the Los Angeles Unified School District (LAUSD). LAUSD 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.[1] More than 76 percent of students are eligible for free/reduced lunches. Increasingly, LAUSD has occupied a visible position in the national conversation surrounding teacher evaluation, and its efforts to train all of its principals is being closely watched by other districts and states.

Population / Participants / Subjects:

LAUSD's new evaluation system is being implemented with all principals and schools in the district in 2012 – 2013. However, consequential decisions linked to evaluation scores will not be implemented until 2013-2014. The observation instrument being used is a modified version of Danielson's *Framework for Teaching* (Danielson & McGreal, 2000) called the Teaching and Learning Framework (TLF). Danielson's original instrument is said to be the most widely used observation protocol in the country. The new instrument, TLF, has been aligned to the California teaching standards.

The size of the observer sample (n=700) allows us to explore the variation in observer thinking and performance across differences in observer characteristics, as well as differences in score quality across school and teacher characteristics.

Intervention / Program / Practice:

The school year 2012-2013 is the first year that the new teacher evaluation system is being adopted district-wide. Principals participated in a training that lasted four days and was designed to help them a.) understand how TLF categorizes and scores teaching practices; b.) accurately score teaching practice using TLF, and c.) take accurate and appropriate notes using TLF. By the end of training, it was expected that observers would be certified to begin observations on the TLF protocol.

To certify as an observer, principals need to demonstrate skill in collecting evidence that is accurate, objective, detailed, and is appropriately used to support scores. Observers must also be able to accurately score teaching practice at acceptable levels, when judged against master observers' scores. Certification status is broken down into four categories of proficiency. If a principal scores in the lowest category, they are not allowed to perform observations.

Research Design and Data Collection:

This study is designed to investigate observer thinking and performance as it occurs in practice, and our current analyses draw on quantitative performance and perception data that comes from more than 700 administrators trained by the district during year 1 of the study. The successful implementation of an observation system such as TLF is largely dependent on the quality of training provided to administrators. At the same time, training success also depends on administrators' willingness to engage in the reform effort. Thus, in this study we explore whether administrators' beliefs prior to training (i.e., their expectations of the training, their beliefs about the uses and usefulness of teacher evaluation, their views of effective teaching, and their feelings of the manageability of their job) are predictive of their satisfaction with training and of their certification success.

[1] Seventy-three percent of students are Hispanic, ten percent are Black, nine percent are White/Non-Hispanic, and six percent are Asian/Pacific Islander. Thirty-one percent of students are English language learners. 32% of the teachers in LAUSD are Hispanic, and 41% are White/Non-Hispanic.

All administrators were given three online surveys: one before, during and after training. For the current study, we draw on the pre-training survey for participants' beliefs and the post-training survey for their satisfaction with the training they receive. We have developed three indicators of training satisfaction: a) participants' self-assessment of learning in training, b) their confidence in their ability to reliably and accurately conduct observations, and c) their assessment of whether the training met their expectations. All survey items included in our analyses are summarized in Table 1.

In addition to the questionnaire data, our study will draw on administrators' certification results, as described in the above section. At present, the district is still processing the certification results; however, these data will be incorporated into the final analyses presented at the Spring 2013 SREE meeting.

Data Analysis:

Research Question 1. Our first research question predicts post-training survey responses related to satisfaction, using pre-training survey data. While we recognize the shortcomings of drawing inferences based on self-reported assessments of training quality, we present these findings to provide context for the implementation of a rigorous observation protocol. For the conference, we will be able to make comparisons between administrators' perceptions of training (i.e., the post-training survey outcomes) and their actual training performance (i.e., the certification results).

In predicting observer satisfaction with training, it was necessary to address the potential variation in training experiences by training site (LAUSD offered four possible sites). To do so, we use hierarchical linear modeling (HLM), treating satisfaction with training as a function of administrators' job characteristics and their perception of a) their job manageability, b) expectations for training, c) beliefs about the importance of teacher evaluation data, and d) whether they believe that instruction should be student-centered.[2] The random intercept HLM model used in our analyses is summarized as follows:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_j$$

where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $u_j \sim N(0, \sigma_u^2)$. Y_{ij} represents the outcome variable for individual administrator *i*, in training site *j*. The term X_{ij} represents the vector of administrator characteristics, beliefs, and expectations. The model is repeated across each of the three indicators of training satisfaction. All continuous level-1 variables are centered by training site in order to ease interpretation of regression coefficients.

Research Question 2. To address Research Question #2, which uses certification status as an outcome, we will draw on the same set of predictor variables listed in Research Question #1; a similar multilevel logit model will be used to predict certification success. However, we will also include post-survey responses to examine whether participants' beliefs after completing training are predictive of their success at certifying. In addition, we will make comparisons across Research Questions #1 and #2 to identify if any predictors prior to training that appear to impact both outcomes.

[2] We use this variable because TLF adopts a perspective which places an emphasis on the important role of students in constructing their knowledge in the classroom. Previous research on another observation protocol suggests this type of view is predictive of certification status (Cash, et al., 2012).

Research Question #3. Previous work suggests that observers have an easier time agreeing with master observers and one another on the classroom organization types of components (i.e., the degree to which students are busy, on task, well behaved, etc.), as compared with scoring instructional and emotional support aspects of classroom interactions (Gitomer, et al., 2012; Bell, et al., 2012). Once certification data become available, we will investigate the specific components observers struggle with and conduct qualitative analyses of the observers' notes in order to better understand what might account for observers' challenges on those components.

Findings / Results:

Results from Research Question #1 are presented in Table 2. Models are run separately for each of the three indicators of training satisfaction: 1) participants' self-assessment of learning in training, 2) their confidence in their ability to reliably and accurately conduct observations, and 3) their assessment of whether the training met their expectations. Across each of the three models, the characteristics of administrators' jobs (e.g., job title, instructional level, years of experience) that we included do not appear to predict training satisfaction. Perhaps not surprisingly, the largest predictor of training satisfaction was administrators' belief in the importance of teacher evaluation for bringing about positive outcomes for teachers and schools. This variable was positive and significant. It also appears, for two of the indicators of job satisfaction, that one's perception of job manageability was an important predictor. We hypothesize that administrators' ability to learn the training material and certify successfully is in part dependent on the time they have available. This is especially true if they are already struggling to manage their responsibilities prior to taking on the new observation duties. Lastly, there is a positive association between observers' initial expectations for training and the degree to which those expectations were met by the training.

Please note that these results are incomplete, and the final paper presented at SREE will include results for Research Questions #2 and #3; i.e., with certification success as an outcome.

Conclusions:

To improve the quality of instruction in schools, districts across the country are investing great hopes and resources in the implementation of teacher evaluation systems. Because observation offers a direct measure of instruction and it can point to areas for teacher improvement, additional research on how to implement observation protocols at scale is imperative and will be highly useful to school districts and other education stakeholders.

The data in this presentation represent one area of UCAST's investigation into the implementation of observation protocols where ratings will be consequential. These data are descriptive and do not support causal inferences about the relationships under study. Knowing which factors predict certification and training satisfaction can guide the design of training and implementation to better use limited training resources and increase the fairness and reliability of observational ratings. The same is true for knowing which components of an observation protocol are particularly challenging.

With so much public attention being given to the events unfolding in LAUSD, the district's choices regarding their teacher evaluation system could serve as a bellwether for other districts in California and across the U.S. We can imagine few districts in the U.S. that could provide us with such an important window onto the challenges and promises of implementing the kind of teacher evaluation system that many districts across the country are adopting.

Appendix A. References

Aaronson, D., Barrow, L., & Sander, W. (2003). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-136.

Bell, C.A., Gitomer, D.H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 1-26.

Casabianca, J., McCaffrey, D., Gitomer, D. H., Bell, C. A., & Hamre, B. H. (2012). Effect of observation mode on measures of secondary mathematics teaching. *Unpublished manuscript.*

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

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). How and why do teacher credentials matter for student achievement? Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching*. (2nd Ed.) Alexandria, VA: Association for Supervision and Curriculum Development.

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

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B., & Pianta, R. (2012). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Unpublished manuscript.*

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review Papers and Proceedings*, 247-252.

Appendix B. Tables and Figures

Not included in page count.

Table 1

Descriptive Statistics of Administrator Sample

Variable	Mean	SD
Administrator characteristics [a]		
Principal (n=399)	0.75	0.43
Assistant Principal (n=110)	0.21	0.41
Years of Experience	6.58	5.73
Elementary School (n=307)	0.58	0.49
Middle School (n=75)	0.14	0.35
High School (n=101)	0.19	0.39
Pre-survey responses (Utility of evaluation data source for improving instruction) [b]		
Observations	3.59	0.59
Student Growth Data	3.01	0.75
Teacher Self-Assessment	3.16	0.77
Student Surveys	2.67	0.85
Pre-survey responses (Utility of evaluation data for various purposes) [b c]		
Identifying or rewarding strong teachers	2.94	0.85
Teacher improvement/development	3.33	0.73
Teacher dismissal	2.86	0.94
School improvement	3.26	0.77
Creating a common vision of excellent teaching	3.33	0.76
Pre-survey responses		
Job manageability [d]	3.27	0.54
High expectations for training [e]	3.65	0.46
Place a value on student-centered instruction [e]	3.81	0.30
Post-survey responses		
Confidence in one's ability to conduct observations reliably [d]	3.36	0.46
Training met ones' expectations [e]	3.41	0.62
Degree to which the administrator learned from training [e]	2.87	0.56

Notes:

[a] With the exception of years of teaching, all are dichotomous variable where 0 = no and 1 = yes

[b] Responses ranged from 1 = Not at all useful to 4 = Highly useful

[c] The individual items listed were included as a composite variable in the analyses

[d] Responses ranged from 1 = Disagree to 4 = Agree

[e] Responses ranged from 1 = Not all to 4 = A lot

Table 2

Predicting Training Satisfaction Using Administrator Pre-Training Survey Responses

	Assessment of learning in training	Training met expectations	Confidence in ability to conduct observations
Admin. Characteristics			
Principal	-0.052	0.082	0.034
	(0.077)	(0.090)	(0.068)
Elementary school	-0.056	-0.063	0.043
	(0.074)	(0.086)	(0.066)
Middle school	-0.015	-0.086	0.007
	(0.090)	(0.104)	(0.078)
Admin. Pre-Training Survey Responses [a]			
Importance of student growth data in teacher eval.	0.080*	0.034	0.022
	(0.041)	(0.048)	(0.036)
Perception of job manageability	0.089*	0.007	0.138***
	(0.051)	(0.059)	(0.045)
High expectations for observation training	0.103	0.234**	0.024
	(0.079)	(0.092)	(0.071)
		0.167***	0.114**
Importance of teacher evaluation data	0.175***		
	(0.052)	(0.060)	(0.047)
Belief in student centered instruction	-0.082	0.056	0.001
	(0.087)	(0.101)	(0.078)
Constant	2.970***	3.411***	3.319***
	(0.080)	(0.087)	(0.067)
Observations	346	346	333
Number of groups	4	4	4

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

[a] Level-1 variables are group-mean centered; the district's four training sites serve as the Level-2 grouping variable