University of Virginia CASTL Curry School of Education

Center for Advanced Study of Teaching and Learning

Research Brief

Large-Scale Rater Calibration for an Observational Instrument[1]

Anne H. Cash, Bridget K. Hamre, Robert C. Pianta, and Sonya S. Meyers

Observational assessment is used to study program and teacher effectiveness across large numbers of classrooms, but training a workforce of raters who can assign reliable scores when observations are used in large-scale contexts can be challenging and expensive. This study reports on the success of rater calibration across 2,093 raters trained by the Office of Head Start on the Classroom Assessment Scoring System™.

Observation is a form of assessment increasingly being used to study children and teachers in school settings. Observational assessments of teacher performance and classroom interactions in early childhood and K-12 settings have been used as a part of research on quality and effectiveness. For example, observation measures were included in several large-scale research studies of early childhood education settings, including the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development and two studies at the National Center for Early Development and Learning. Also, the Bill and Melinda Gates Foundation recently designated $45 million for research on teacher effectiveness in K-12 settings. A portion of this funding is dedicated to videotaping and observing 3,000 teachers using each of five different observation protocols.

More recently, states have been using observational assessments to evaluate teacher quality in response to the federal Race to the Top and Early Learning Challenge initiatives. For example, observational measures are often included as part of states' Quality Rating and Improvement Systems for early childhood education settings.

Two significant concerns exist for the people who coordinate large-scale observational assessments.

1. Can staff from varying backgrounds be trained to perceive classrooms in the same way? The level of training required to establish acceptable inter-rater reliability on observational measures varies depending on characteristics of the observation and of the observer and can require intensive resources in terms of time and money.

2. What are the characteristics of people who can assign reliable scores? For all of the effort required to coordinate observational assessments in large-scale contexts, and for the impact of decisions resulting from these efforts, establishing a workforce of calibrated raters is a crucial first step.

Standardized observation tools provide evaluators, raters, teachers, and administrators with a common metric for assessing and improving quality. The dilemma, however, is summarized as follows: When implementing large-scale projects, evaluators want rater training to take just a few days with limited follow-up, but they lack information on the feasibility of this preference. They often want to use their own staff for leading observer trainings, but whether staff can efficiently become well-versed in the tool is unclear. Finally, evaluators want to be able to hire people who can calibrate to the tool right away without lots of expensive follow-up. Few evidence-based guidelines are available on identifying the best candidates from a pool of people with diverse experiences and beliefs about teaching.

Because most observation tools have been developed and used in smaller projects requiring fewer than 100 raters, opportunities to examine issues of rater calibration relevant to large-scale contexts have been limited.

The Study

The Improving Head Start for School Readiness Act of 2007 required the Office of Head Start to include a valid and reliable observational tool for assessing program quality. A nationwide effort was undertaken to train practitioners on an observational measure of teacher-child interactions, the Classroom Assessment Scoring System™ (CLASS). Researchers from the Center for Advanced Study of Teaching and Learning, where the CLASS was developed, presented CLASS trainings as opportunities to build staff capacity to

[1]This research brief is based on the following published study: Cash, A. H., Hamre, B.K., Pianta, R.C., & Myers, S.S. (2012). Rater calibration when observational assessmentoccurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542
This published study can be purchased at: http://www.sciencedirect.com/science/article/pii/S0885200611000974

Research Brief © 2013 Rector and Visitors of the University of Virginia. For more information please visit www.curry.virginia.edu/castl or contact castl@virginia.edu

The Classroom Assessment Scoring System™

The CLASS™ is an observational instrument developed at the University of Virginia to assess classroom quality in PK-12 classrooms. It describes multiple dimensions of teaching that are linked by research to students' positive social and academic and development and has been validated in thousands of classrooms.

The CLASS™ instrument addresses three broad domains of effective teacher-student interactions: Emotional Support, Classroom Organization, and Instruction Support. It can be used to reliably assess classroom quality for research and program evaluation and also provides a tool to help new and experienced teachers become more effective.

For more information about the CLASS™, see ***http://curry.virginia.edu/research/labs/class***

assess and improve classroom quality in their programs.

Use of the CLASS as an observational instrument or professional development tool by Head Start programs was strictly voluntary. However, this project presented a unique opportunity to study raters and was one of few large-scale efforts to systematically gather data on practitioners' calibration to a research-based observation tool. Scores from their calibration assessments provided evidence of the success of a scaled-up approach to training raters.

Of 2,093 participants who completed the calibration assessment, 704 also elected to complete a brief survey at the beginning of the training session, in which they reported on demographic characteristics, job responsibilities, and beliefs about teaching.

Among those who completed the survey, 13% had an associate's degree or less, 48% had a bachelor's degree, and 37% had a master's degree or higher. They reported an average of 9 years of experience supervising or mentoring teachers, and 43% had been in their current position with Head Start for 1–5 years.

CLASS rater trainings were led by 25 trainers who were also Head Start Training and Technical Assistance specialists responsible for working directly with Head Start programs to help them meet monitoring and performance standards.

The trainings were each three days long. In the first two days, the CLASS structure and coding protocol were introduced and trainees practiced coding five 20-min video segments of real preschool classrooms. There was time for trainees to ask questions and engage in discussion to further develop their understanding of the CLASS and the coding process. The calibration assessment took place on the third day. For this assessment, trainees watched three 20-min video segments and spent 20 minutes coding each one. Their codes were compared to a set of master codes previously established by three expert CLASS coders.

Results

Overall, this study provided evidence that it is possible to train large numbers of raters to calibrate to an observation tool through 2-day training sessions led by the evaluator's own staff, when that staff has been trained as trainers on the tool.

Of Head Start staff trained on the CLASS tool, 71% passed the calibration assessment on their first attempt. This finding indicates that creating a workforce of calibrated raters is possible, but not all raters will pass the first time.

Some raters found it easier to calibrate to the CLASS than others. Raters who believed intentional teaching practices are important were more closely calibrated with the master codes. When the beliefs of a group of raters were more adult centered than child centered on average, raters were less calibrated overall, and particularly for the dimension Regard for Student Perspectives.

Also, some components of the instrument may have been more difficult for raters to calibrate to. Raters appeared to assign scores higher than the master codes in the CLASS dimensions of Concept Development, Quality of Feedback, and Language Modeling (from the Instructional Support domain) and the dimension Regard for Student Perspectives. They weighed certain examples of teacher-child interactions too heavily in assessing their quantity and quality. This rating behavior is consistent with anecdotal reports that the Instructional Support domain is difficult to teach and to learn.

Possibly, these dimensions of the CLASS instrument reflect a way of thinking and talking about teaching that are different from common understanding and require more of a shift in beliefs and knowledge for newly trained raters (in this case raters who are also practitioners) to calibrate. When rater beliefs are misaligned with the theoretical foundation of an observational assessment, calibration could be problematic.

The findings that rater beliefs are important, particularly for certain dimensions, have implications both for how raters are trained and who should be trained to do observational assessments at scale.

Future implementers of large-scale observational assessments can be reassured that it is possible to train large numbers of raters to calibrate to the CLASS in a short period of time. When hiring raters, evaluators should pay special attention to rater beliefs and particularly the ways in which rater beliefs are or are not aligned with the CLASS. Also, if specific components of the observation tool are controversial or offer a way of thinking that is less than common knowledge among the selected raters, evaluators should allow for sufficient time during trainings to expose and disperse rater bias associated with those components.

This research brief should be cited as: Cash, A. H., Hamre, B.K., Pianta, R.C., & Myers, S.S. (2012). Large-scale rater calibration for an observational instrument. Retrieved from: http://curry.virginia.edu/uploads/resourceLibrary/Research_Brief_-_Rater_calibration.pdf

Research Brief © 2013 Rector and Visitors of the University of Virginia. For more information please visit www.curry.virginia.edu/castl or contact castl@virginia.edu