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™.
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.