Working Paper:

Measuring the Quality of Teacher-Child Interactions at Scale: The Implications of Using Local Practitioners to Conduct Classroom Observations

Virginia E. Vitiello[1], Daphna Bassok[1], Bridget K. Hamre[1], Daniel Player[1] & Amanda P. Williford[1]

Use of observational measures to monitor preschool quality is growing rapidly. This paper examined the degree of agreement between local and research rater teams using an observational measure of preschool classroom quality, and the extent to which ratings predicted gains in children's literacy, math, and self-regulation skills. Both rating teams observed 85 classrooms using the Pre-K CLASS and 820 children (average age = 52.6 months, SD = 3.6 months) were directly assessed in the fall and spring. Results indicated moderate correlations between local and research teams' scores, ranging from r = .19 to .40. Both teams' scores significantly predicted child gains, although patterns of association differed. Results are discussed in the context of policies that require observational measures at scale.

[1]University of Virginia

Updated November 2016

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

Acknowledgements: This research was supported by a grant from the Institute of Education Sciences (R305A140069). Opinions reflect those of the authors and do not necessarily reflect those of the granting agency. We thank the Louisiana Department of Education for their willingness to share data for this project, and the children, teachers, and families who generously agreed to participate in this study.

MEASURING THE QUALITY OF TEACHER-CHILD INTERACTIONS AT SCALE: THE IMPLICATIONS OF USING LOCAL PRACTITIONERS TO CONDUCT CLASSROOM OBSERVATIONS

Virginia E. Vitiello, Daphna Bassok, Bridget K. Hamre, Daniel Player & Amanda P. Williford

Early childhood programs can yield large short and long-term benefits (Bassok & Loeb, 2015). However, the quality of early childhood programs in the United States is highly variable. To address this problem, Quality Rating and Improvement Systems (QRIS), which are accountability systems aimed at accurately measuring program quality and incentivizing improvements, have been introduced nearly nationwide.

While the logic of these quality efforts is compelling it is not yet clear whether QRIS will lead to meaningful improvements in program quality. One concern is that the rapid design and roll-out of states' QRIS systems has outpaced the research base around measuring and improving quality. In order to lead to quality improvements, QRIS must accurately measure the features of program quality that affect child learning. However, many state systems provide ratings that are unrelated to children's developmental outcomes (Sabol, Hong, Pianta, & Burchinal, 2013; Sabol & Pianta, 2014).

Observed measures of teacher-child interactions have shown particular promise in predicting positive child outcomes and providing teachers with specific, behavioral feedback to improve quality and thus are a focus for observation in many QRIS (Hamre, 2014). These teacher-child interaction measures are arguably the best available measure of program quality, as they are more predictive of child outcomes than other available measures such as teacher education or class size (Mashburn et al., 2008). For this reason, the vast majority of states currently include classroom observation as a component of their rating systems (The BUILD Initiative & Child Trends, 2015). However, research studies documenting positive associations between classroom observations and children's learning typically rely on *researcher-collected* observations. For practical and cost considerations, some QRIS are turning to local, non- researcher observers, who may not have reliability supports in place and who may also be less objective compared to external observers. Increasingly, these observations are being used in high stakes situations where the reliability of these data is of paramount importance. There is currently very limited research examining the reliability and validity of these measures at scale, or whether community-based raters can reliably assess local classrooms. This is the gap we fill in the current paper.

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Observational Ratings of Teaching

The Classroom Assessment Scoring System (CLASS: Pianta, LaParo, & Hamre, 2007) is the most commonly used observational measure of teacher-child interactions. Research using CLASS demonstrates that when teachers offer warm, supportive and responsive interactions, children develop stronger social and emotional skills (e.g., Johnson, Seidenfeld, Izard, & Kobak, 2012). Children in classrooms with strong behavior management and classroom organization demonstrate stronger growth in self-regulation (Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). Further, teachers' daily provision of cognitively stimulating instruction and conversation appears to be a critical ingredient in fostering academic learning (e.g., Howes et al., 2008).

Notably, however, these studies use researcher-trained raters who typically receive substantial supports to maintain their reliability, including periodic re-calibration sessions and opportunities to double-code a subset of observations and debrief on those double codes. As yet, there has been no research to determine whether observations collected by local coders show the same links to child outcomes that have been demonstrated across many research studies. There is reason for concern given research from K-12 settings suggesting significant rater effects (Gargani & Strong, 2014; Mashburn et al., 2014) and that local raters (e.g. principals) may inflate scores (Ho & Kane, 2013). The fairness and ultimate impact of QRIS depends upon an assumption that programs will receive similar scores regardless of who is completing the rating.

Quality Rating and Improvement System in Louisiana

As part of Louisiana's 2012 Education Reforms, the state passed the Early Childhood Education Act (known as Act 3), an effort to overhaul the state's fragmented early childhood landscape into a cohesive system. When fully implemented, Act 3 requires Louisiana to have a unified early childhood accountability system in which all publicly funded child care, Head Start, and pre-k programs must participate. This system will have all the defining components of a QRIS, including quality ratings for programs, financial incentives, supports for improvement, and public information campaigns for parents. However, whereas in most states observational measures of quality are one of a host of quality measures included in the QRIS, in Louisiana, CLASS observations are the *only* quality measure that will be used to calculate program ratings at this time. When fully-implemented, publicly-funded programs that do not participate or fail to meet a minimum standard can lose licensing or funding.

CLASS observations are coordinated and collected separately in each Community Network, which is typically a group of all the publicly-funded early childhood programs in a particular parish that is coordinated by a single lead agency (e.g. a school district). The state funds lead agencies to conduct CLASS observations and requires that all observers attend a CLASS training and pass the certification test. Beyond that, local networks have substantial flexibility with respect to the ways in which they provide ongoing support to raters.

Current Study

The current study adds to our knowledge about the role of raters in establishing a fair, reliable, and valid assessment of quality, first by examining the extent to which local and research-trained teams of raters come to similar conclusions about the quality of classrooms and programs using the CLASS, and then by comparing whether these teams' observation scores predict children's directly-assessed achievement gains in similar ways. There is a strong need for more research to guide decision-making in QRIS, including research on the implementation of observational rating systems. This study fills that gap.

Methods

Participants

Primary data collection included 90 classrooms in five Louisiana parishes. At the end of the academic year, the state department of education provided the research team with CLASS data from the local coding teams. Local coders visited 85 of the 90 classrooms observed by the research coding team; we limit the current analysis to those 85 classrooms. The classrooms were located in public schools (45.9%), private childcare centers (12.9%), Head Start centers (18.8%), charter schools (11.8%), and private schools and centers receiving state subsidies (10.6%). All children who were four years of age and had no IEP (except for IEPs related to language delays) were eligible to participate. The child sample included 820 predominantly low-income four-year-old children who were assessed in both the fall and spring. Teacher and child demographic characteristics are presented in Table 1.

Measures

Classrooms were observed using the CLASS, an observation of teacher-child interactions that assesses effective interactions across ten dimensions divided into three broad domains,

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Emotional Support, Classroom Organization, and Instructional Support (Pianta, et al., 2007). Both research team and local raters conducted CLASS observations in multiple 30 minute cycles that include 20 minutes to observe and record classroom interactions and 10 minutes to code the CLASS dimensions; codes from each cycle are averaged together to arrive at a single set of classroom scores. Dimensions are coded on a seven-point scale with detailed behavioral descriptors of interactions at the low (1-2), mid (3-5), and high (6-7) ranges of effectiveness. Fifteen percent of observations conducted by the research coding team were double-coded by two data collectors. Intraclass correlations indicated high levels of agreement between research raters (ICC range from 812 to .902). Internal consistency estimates were strong across both rater teams, with Cronbach's alphas ranging from .77 to .96.

Children were directly assessed using tests of language, literacy, math, and executive functions. The Peabody Picture Vocabulary Test-4th edition (PPVT-IV) was used to measure children's receptive vocabulary skills (Dunn & Dunn, 2007). Three subtests of the Woodcock – Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001) were used to assess children's expressive vocabulary (Picture Vocabulary) and math skills (Applied Problems and Quantitative Concepts). The Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, Rashotte, 2007) Phonological Awareness and Print Knowledge subtests were used to assess children's literacy skills. Executive functions were assessed using the Pencil Tap test of inhibitory control (Smith-Donald, Raver, Hayes, & Richardson, 2007) and the Head Toes Knees Shoulders task (HTKS; Ponitz et al., 2008) measure of inhibitory control, working memory, and attention. All have been used extensively in research and validated for use with low income preschool children.

Procedure

In collaboration with the state department of education, the research team selected five Louisiana parishes to participate in the study that captured the geographic and demographic diversity of the state. Private and state-funded programs were eligible if they were participants in the state's early child care network, an initiative designed to start building a unified, statewide approach to early childhood education. Ninety preschool programs serving four-year-olds were randomly selected stratified by program type, from a list of eligible programs provided by the state. The acceptance rate, among invited programs, was 84-100% by parish.

Within each program, all teachers of classrooms serving primarily four-year-olds and typically developing children were randomly ordered and the first teacher from each program was

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

contacted. Six teachers declined to participate or were later found to be ineligible, so the next eligible teacher on the list was contacted. If there were no other eligible teachers at the program, the program was dropped and another program was contacted as a replacement. Four teachers left their classrooms during the year and were replaced with the teacher who took over teaching responsibilities for that classroom.

All research and local raters completed the Teachstone CLASS Observation Training and passed the reliability test with a minimum score of 80% agreement within one point of master codes. The research coding team was required to complete one day of live coding practice with an experienced coder, two calibration sessions during data collection, and frequent double-coding and debrief sessions with fellow raters. The research coding team did not personally know the teachers that they observed.

Parishes varied in their approaches to supporting local raters. Three parishes did some proportion of double coding, and four of the five reported that raters sometimes, often, or almost always knew the teachers they observed. The research coding team visited each classroom an average of 3.94 times (*SD* = .28); local raters visited an average of 1.47 times (*SD* = .59). Both teams coded four cycles per visit.

Direct assessments were completed by the research team in the fall and spring. Children were assessed individually in a single session lasting approximately 45 minutes in a quiet location outside of the classroom, as free from distractions as possible.

Analytic approach. Due to differences in the number and timing of observations between research team and local raters, this study relies on one randomly selected day of ratings from each classroom for each of the local and research coding teams; CLASS scores for each team therefore represent scores derived from four cycles of observation collected on a single day. Louisiana's QRIS currently uses the following cut points for the CLASS total score to categorize programs: Unsatisfactory: 1 – 2.99; Approaching Proficient: 3 – 4.49; Proficient: 4.50 – 5.99; Excellent: 6 – 7. We use the same cut points to compare alignment across raters. To analyze the child assessment data, assessment scores were converted into z-scores centered at the fall mean and then were averaged together to form three composites: math, literacy (including language scores), and executive functions. The three composites were also averaged together to create an average child achievement score. Separate regressions were run using the CLASS scores to predict spring achievement controlling for fall scores and classroom fixed effects.

Results

Means. Domain and total score means from each coding team are presented in Table 2. Paired samples *t*-tests suggest that Total CLASS scores were marginally higher among local raters ($p = .058$). Local raters' Instructional Support scores were roughly three quarters of a point higher than the research team's ($p < .001$). For Classroom Organization, the research team coded classrooms marginally higher than the local raters ($p = .055$).

Correlations. Domain and total score correlations across teams are presented in Table 3. Within each coding team domain scores were moderately to highly correlated, ranging from $r = .48$ to $r = .68$ for the research team and $r = .65$ to $r = .82$ for the local raters. Across coding teams, domain scores were significantly correlated with one exception: the correlation between the research team's Emotional Support and the local raters' Instructional Support did not reach significance ($r = .19$, $p = .075$). The remaining correlations tended to be in the low to moderate range, from $r = .24$ to $r = .38$.

Rating categories. CLASS total scores from both rater teams were divided into quality categories using the Louisiana QRIS cut points (Table 4). Results indicated moderate agreement across raters, with 55.3% of classrooms placed into the same category by both teams. In 25% of classrooms the local team rated the classroom in a higher category than the research team, the reverse was the case in 20% of classrooms. There was no agreement on membership in the Unsatisfactory or Excellent categories. Of particular note, local coders placed 9 classrooms in the Excellent category and the research team placed 2, but none was rated excellent by both.

Prediction to child gains. Results from regression analyses are presented in Table 5. Both teams' CLASS scores showed significant prediction to child gains. However, the pattern of significance differed. For the local raters, all CLASS scores were significantly associated with children's literacy gains. Instructional Support, Classroom Organization, and the total score were also associated with math gains. None of the local raters' CLASS scores were associated with gains in executive functions. For the research team, math, literacy, and executive functions were all significantly associated with Emotional Support, Classroom Organization, and the total score. Instructional Support scores were not associated with any of the child outcomes. Both teams showed significant associations between the CLASS total score and children's average achievement gains. Associations between CLASS and child outcomes were generally stronger for research team ratings; in seven cases, these differences in the magnitude of association were statistically significant.

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Discussion

With the rapid expansion of QRIS nationwide, and the increasing use of observational measures within them, there is a need for research that can guide decision-making in ways that help ensure the fairness and reliability of data (Lahti et al., 2015). The use of local observers offers several notable potential benefits (e.g., saving money, gaining local buy-in), but their use may also create unintended consequences producing biased or unreliable scores. In comparing local coders' ratings to ratings conducted by a research team, the current study evaluated the reliability and validity of local ratings for use in a QRIS. The results provide a mixed picture.

The study suggests that classroom observations of teacher-child interactions do demarcate important elements of children's classroom experiences. Classrooms scoring higher on the CLASS, as assessed by either local or research raters, had children that made greater gains across domains of development. Thus classrooms that receive higher ratings in the Louisiana QRIS appear to be supporting better outcomes for students, a claim that not all QRIS can make (Sabol et al., 2013).

Average CLASS scores for classrooms were similar for two of the three domains - Emotional Support and Classroom Organization. Correlations between the CLASS total score given by research and local raters were moderate and roughly half of the classrooms were placed into the same "proficiency" category by both teams. These patterns are encouraging given that local raters demonstrated fair levels of overlap with research raters, who typically had more intensive training and support.

There were, however, areas in which ratings by local and research teams diverged. From a policy perspective, the most notable divergence was that many classrooms were put into different QRIS categories. There was no concordance in the highest and lowest rating categories. Although assigning these "extreme" categories were uncommon for both raters, classrooms were 4.5 times as likely to be placed into the "Excellent" category by local raters than by the research raters, driven in large part by the significantly higher Instructional Support ratings they assigned. The potential "inflation" of scores by local raters is in line with prior research on teacher evaluation in K-12 (Ho & Kane, 2013).

There is some evidence that local teams were assigning ratings that tapped into *global* elements of quality interactions rather than distinguishing between interactions in the specific domains. The local coders' domain scores were more strongly correlated with each other than were the research coders' domain scores, suggesting lower discrimination between the domains. As noted

earlier, the strongest association across the two teams was on the CLASS total score, suggesting that the total score may provide the most reliable estimate of interaction quality.

Some states and Head Start have chosen to place an emphasis on each separate domain of CLASS scores within their QRIS by requiring programs to meet separate thresholds on Emotional Support, Classroom Organization, and Instructional Support. This is based, in part, on prior work suggesting domain specificity in the association between CLASS scores and outcomes – with Instructional Support typically predicting gains in academic learning, Classroom Organization predicting gains in EF, and Emotional Support predicting social development (Hamre, 2014). This study did not replicate those findings. The specific patterns of prediction from CLASS rating to children's outcomes are difficult to interpret and will require further study. However, the current results do provide support for the strategy used by Louisiana to rely on the CLASS total score for program ratings. For both local and research raters the total score had the most consistent and strongest associations with children's learning and development across domains. Though the provision of the more refined scores for the CLASS domains (i.e., Emotional Support) or dimensions (e.g. Teacher Sensitivity) may be useful for professional development (Pianta et al., 2014), the total score may be a better measure to use in high stakes settings.

Several caveats about this work are worth noting. These data were collected during Louisiana's pilot year. First, the CLASS observations were not tied to incentives or consequences in that year, though they soon will. The pressure to assign higher codes may be more pronounced when the observations are high stakes. Second, completing CLASS observations was a new responsibility for lead agencies and there were no requirements that local community networks ensure the reliability of observations beyond requiring the initial certification. It is worth noting that Louisiana has put into place a number of measures in the years since the pilot to help ensure reliability. For instance, the state now sends, independent, "third-party" observers commissioned by the state to conduct observations in classrooms at every site, and uses those third-party observations when there is a large divergence with the local ratings. The state has recently developed stringent guidelines for ensuring the reliability of these third-party observers. These may be effective at preventing coding drift over time (Karoly et al., 2013).

An additional set of limitations relate to the way data were collected for this study. First the sample size is relatively small which limits the extent to which the results can be generalized; future studies will need to evaluate the reliability and validity of local ratings using a more extensive sample across the state. Second, reliability and validity were assessed here at the *classroom* level and only in

preschool classrooms, when most QRIS (including Louisiana's) assign ratings at the *program* level and include ratings for infant and toddler classrooms as well. In addition, observations were not conducted on the same day, thus adding additional error into our assessment of alignment between local and research raters (Casabianca et al., 2013).

These findings suggest several recommendations for policy makers and researchers. Given the moderate evidence for reliability and validity across coding teams, we urge caution in incorporating local observation scores into formal QRIS ratings. There is evidence supporting the use of local observers, but strong data quality procedures should be in place to ensure the best possible data, including reliability testing procedures, calibration opportunities during data collection, and frequent checks on the data to make sure scores from different teams are well aligned. Particular attention should be paid to ways to help raters understand and score Instructional Support.

For researchers, this work highlights the need for research that can inform QRIS development and decision-making. There are many new practices being included in these systems that have not been adequately studied, and there are significant policy implications for this work. Specific to classroom observations, future work may focus on understanding the factors that affect reliability such as the timing of observations, the number of classroom visits, the calibration processes in place, and how raters are assigned to classrooms.

References

Bassok, D., & Loeb, S. (2015). Early childhood and the achievement gap. In H. F. Ladd & M. Goertz (Eds.), *Handbook of Research in Education Finance and Policy (2nd ed.)*.

Boller, K., Paulsell, D., Del Grosso, P., Blair, R., Lundquist, E., Kassow, D. Z., ... & Raikes, A. (2015). Impacts of a child care quality rating and improvement system on child care quality. *Early Childhood Research Quarterly*, *30*, 306-315.

Bryant, D. (2010). Observational Measures of Quality in Center-Based Early Care and Education Programs, OPRE Research-to-Policy, Research-to-Practice Brief OPRE 2011-10c. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. *Quality Measurement in Early Childhood Settings*. Baltimore, MD: Paul H.Brookes.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25*, 166-176.

Burchinal, M., Vernon-Feagans, L., Vitiello, V., & Greenberg, M. (2014). Thresholds in the association between child care quality and child outcomes in rural preschool children. *Early Childhood Research Quarterly, 29*, 41-51.

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.

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*, 757-783.

Dunn, L. M. & Dunn, D. M. (2013). *PPVT-4 Technical Report.* Minneapolis, MN: Pearson Assessments.

Dunn, L. M. & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test – 4th Edition*. Minneapolis, MN: Pearson Assessments.

Gargani, J., & Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. *Journal of Teacher Education*, 0022487114542519.

Goffin, S. G., & Barnett, W. S. (2015). Assessing QRIS as a change agent. *Early Childhood Research Quarterly*, *30*, 179-182.

Gosse, C. S., McGinty, A. S., Mashburn, A. J., Hoffman, L. M., & Pianta, R. C. (2014). The role of relational and instructional classroom supports in the language development of at-risk preschoolers. *Early Education & Development*, *25*(1), 110–133. doi: 10.1080/10409289.2013.778567

Grossman, S. R., Seidenfeld, A. M., Izard, C. E., & Kobak, R. (2012). Can classroom emotional support enhance prosocial development among children with depressed caregivers? *Early Childhood Research Quarterly*, *28*(2), 282–290. doi: 10.1016/j.ecresq.2012.07.003

Hindman, A. H., & Wasik, B. A. (2013). Vocabulary learning in Head Start: Nature and extent of classroom instruction and its contributions to children's learning. *Journal of School Psychology*. *51*(3), 387-405. doi: 10.1016/j.jsp.2013.01.001

Ho, A. D., & Kane, T. J. (2013). *The Reliability of Classroom Observations by School Personnel. Research Paper*. Seattle, WA: MET Project, Bill & Melinda Gates Foundation.

Hong, S. L. S., Howes, C., Marcella, J., Zucker, E., & Huang, Y. (2015). Quality Rating and Improvement Systems: Validation of a local implementation in LA County and children's school-readiness. *Early Childhood Research Quarterly*, *30*, 227-240.

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., … Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171-1190. doi: 10.1111/cdev.12048

Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly*, *30*, 280-290. doi: 10.1016/j.ecresq.2014.04.005

Le, V. N., Schaack, D. D., & Setodji, C. M. (2015). Identifying baseline and ceiling thresholds within the Qualistar early learning quality rating and improvement system. *Early Childhood Research Quarterly*, *30*, 215-226.

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). Test of Preschool Early Literacy (TOPEL). Austin, TX: Pro-Ed.

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, *15*(2), 146-155.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732-749.

Merritt, E. G., Wanless, S. B., Rimm, S. E., Cameron, C., & Peugh, J. L. (2012). The contribution of teachers' Emotional Support to children's social behaviors and self-regulatory skills in first grade. *School Psychology Review*, *41*(2), 141–159.

Pianta, R. C., DeCoster, J., Cabell, S., Burchinal, M., Hamre, B. K., Downer, J., ... & Howes, C. (2014). Dose–response relations between preschool teachers' exposure to components of professional development and increases in quality of their interactions with children. *Early Childhood Research Quarterly, 29,* 499–508.

Ponitz, C. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. Early Childhood Research Quarterly, 23, 141–158.doi:10.1016/j.ecresq.2007.01.004.

Sabol, T. J., Hong, S. S., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-K programs predict children's learning?. *Science, 341(6148),* 845-846.

Sabol, T. J., & Pianta, R. C. (2014). Do Standard Measures of Preschool Quality Used in Statewide Policy Predict School Readiness?. *Education Finance and Policy*, *9*(2), 116-164.

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. Early Childhood Research Quarterly, 22, 173–187. doi:10.1016/j.ecresq.2007.01.002

Sonnenschein, S., Thompson, J. A., Metzger, S. R., & Baker, L. (2013). The importance of teachers' language and children's vocabulary to early academic skills. *NHSA Dialog*, *16*(4), 107–112.

The Build Initiative & Child Trends. (2015). *A catalog and comparison of Quality Rating and Improvement Systems (QRIS)* [Data System]. Retrieved from http://qriscompendium.org/ on 8/9/2016.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III Tests of Achievement. Itasca, IL: Riverside Publishing.

Table 1. Participant characteristics.

	Percent	Mean	SD
Teachers			
Female	98.4		
Race/Ethnicity			
White	54.1		
Black/African American	41.2		
Hispanic	2.4		
Other ethnicities	2.3		
Education			
Associate's degree	4.7		
Bachelor's degree	41.2		
Bachelor's degree plus additional coursework	22.4		
Master's degree	21.2		
Beyond a Master's	4.7		
Other	4.7		
Years experience		9.5	8.9
Children			
Female	49.6		
Age (months)		52.6	3.6
Race/Ethnicity*			
White	20.8		
Black/African American	70.4		
Hispanic	2.7		
Other ethnicities	6.1		
Family Income*			
$25,000 or less	67.4		
$25,001 – 55,000	23.9		
$55,001 or more	8.6		
Parent Education*			
No high school diploma	14		
Diploma or GED	13		
Some college, no degree	30.9		
Associate's degree	11.3		
Bachelor's degree or higher	13.2		

**Available for children whose parents completed demographic items on a parent questionnaire (n = 568-631).*

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 2. CLASS score means and standard deviations.

	Local Raters				Research Team Raters				
	Mean	Std. Dev	Min	Max	Mean	Std. Dev	Min	Max	
Total	4.98	0.89	2.01	6.94	4.79	0.67	2.95	6.62	†
Emotional Support	5.80	0.77	2.63	7.00	5.79	0.63	4.06	6.94	
Classroom Organization	5.34	0.87	2.33	7.00	5.54	0.80	3.00	6.92	†
Instructional Support	3.79	1.33	1.08	6.83	3.05	0.95	1.33	6.08	***

† $p < .10$; *** $p < .001$

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 3. Correlations between domain and total scores across coding teams.

		Local Raters				Research Team			
		ES	CO	IS	Total	ES	CO	IS	Total
Local Raters	ES	--							
	CO	.82	--						
	IS	.65	.66	--					
	Total	.88	.89	.90	--				
Research Team	ES	.27	.29	.19	.27	--			
	CO	.24	.37	.31	.35	.68	--		
	IS	.30	.34	.25	.33	.57	.48	--	
	Total	.33	.40	.30	.38	.85	.84	.84	--

Note: All correlations are statistically different from zero with $p<.05$ except the correlation between the Research Team Emotional Support and Local Rater Instructional Support which has a $p=.075$. ES = Emotional Support; CO = Classroom Organization; IS = Instructional Support; Total = Total CLASS Score.

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 4. QRIS category frequencies and agreement across raters

		Research Team				
		Unsatisfactory	Approaching Proficient	Proficient	Excellent	Total
Cut points		1 – 2.99	3 – 4.49	4.50 – 5.99	6 - 7	
Local Coders	Unsatisfactory	0 (0%)	3 (3.5%)	0 (0%)	0 (0%)	3 (3.5%)
	Approaching Proficient	1 (1.2%)	8 (9.4%)	12 (14.1%)	0 (0%)	21 (24.7%)
	Proficient	0 (0%)	11 (12.9%)	39 (45.9%)	2 (2.4%)	52 (61.2%)
	Excellent	0 (0%)	1 (1.2%)	8 (9.4%)	0 (0%)	9 (10.6%)
	Total	1 (1.2%)	23 (27.1%)	59 (69.4%)	2 (2.4%)	85 (100%)

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 5. Associations between ratings and child achievement gains.

Local Raters	Math	Literacy	Executive Function	Achievement Average
Emotional Support	0.067	0.098 *	0.059	0.077
	(.052)	(.046)	(.056)	(.048)
Classroom Organization	0.110 *	0.112 *	0.076	0.097 *
	(.052)	(.043)	(.041)	(.043)
Instructional Support	0.110 **	0.081 **	0.052	0.083 **
	(.035)	(.03)	(.028)	(.030)
CLASS Total Score	0.136 **	0.122 **	0.078	0.114 *
	(.049)	(.044)	(.042)	(.044)
Research Team				
Emotional Support	0.210 **	0.122 *	0.235 ***	0.184 **
	(.064)	(.056)	(.057)	(.058)
Classroom Organization	0.253 ***	0.174 ***	0.184 ***	0.209 ***
	(.054)	(.043)	(.042)	(.046)
Instructional Support	0.023	0.005	0.019	0.004
	(.049)	(.037)	(.038)	(.040)
CLASS Total Score	0.199 ***	0.122 *	0.172 **	0.157 **
	(.058)	(.05)	(.054)	(.053)

*Note: *p < .05; **p < .01; ***p < .001.* Shaded cells reflect coefficients that are significantly different for the research team compared to the local raters.

EdPolicyWorks Working Paper Series No. 52. November 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia