Early Childhood Research Quarterly 51 (2020) 285–287

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Measuring and improving quality in early care and education☆

Robert C. Pianta*, Bridget K. Hamre, Tutrang Nguyen

University of Virginia, United States

ARTICLE INFO

Article history:
Received 28 March 2019
Received in revised form
10 September 2019
Accepted 31 October 2019
Available online 8 January 2020

Keywords:
Child care quality
Early childhood education
Measurement

ABSTRACT

For decades, research on the quality of early care and education has helped identify features of those settings that play some role in promoting children's learning and development. We offer this brief commentary as a framework for considering not only the topic and papers in this special issue, but more importantly, as a motivating heuristic for next generation of research and development on the assessment and improvement of measuring quality. First, we present some observations on the contemporary debates about quality, including a discussion of effect sizes, causality, conducting ẗhorse raceäanalyses, and bringing measures to scale. We then argue that making progress towards measuring quality requires the field to think carefully about the focus of assessment, tradeoffs between reliability and validity, the diversity of the classroom and teachers, and scientific and technical advances. Absent these considerations, individually and in terms of their integrative links, discussions or presentations of n̈ew and improvedm̈easures of quality will not help advance the understanding and appropriate use of such instruments.

In this paper, two of us (RCP, BKH) reflect briefly on more than 15 years developing, evaluating, implementing, and scaling a method to observe and improve "quality," defined as teacher-student interaction. Three themes emerged: 1) the role of theory; 2) interpreting links between quality and child outcomes; and 3) lessons learned from working at scale.

1. The role of theory

Every "measure" of quality is a theory, hypotheses about the processes that foster children's development, and how best to gather information. Each time a measure is applied, data confirm or disconfirm these hypotheses. Throughout our work on the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) we tried to be anchored in theory, test hypotheses, with an eye on application.

We defined quality in terms of teacher-student interactions because of evidence from attachment theory and the predictive value of mother–child relationships and interactions in studies of development (Sroufe, 1996). Theory informed basing all CLASS dimensions around properties of "serve and volley" exchanges that required attention to both the teacher's behavior *and* the child's response (Sroufe, 1996). Our theory of classroom processes, the *Teaching Through Interactions* framework (TTI; Pianta & Hamre, 2009), hypothesized a taxonomy that organized definitions of interaction at four levels. Distinct from other measures organized as lists of discrete behaviors (e.g., smiling, attending to a child), the TTI/CLASS is a *latent* structure reflecting developmentally salient processes of adult–child relationship systems. Definitional details within three broad domains cascade into increasingly granular levels of analysis—dimensions, markers, and behavioral indicators.

Theory predicted that this latent structure would apply across all grade levels, content areas, or focus of instruction, which also solved challenges in application (described later). This common latent structure and the value of the interactive properties of teacher–student relationships across grades found empirical support in: studies validating the three-factor latent structure (Hamre et al., 2013) and associations with student outcomes (Allen et al., 2013); analysis of alternative latent organizers using bi-factor models at different grades (Hamre, Hatfield, Pianta, & Jamil, 2014); and intervention studies (Allen et al., 2013).

☆ We gratefully acknowledge the support of our many partners: school district leaders, community programs, teachers, parents and students who participated in the work summarized here. Their enthusiastic cooperation and participation made much of this work possible. We also extend appreciation and recognition to faculty colleagues, staff members, graduate students, and fellows who were part of this program of research, and the educators and policy-makers who have been involved in its application, all made valuable contributions. Current support for this research is provided by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305N160021 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Request reprints from author Pianta at rcp4p@virginia.edu.

* Corresponding author at: University of Virginia, Curry School of Education, P.O. Box 400260, Charlottesville, VA 22904, United States.
E-mail address: rcp4p@virginia.edu (R.C. Pianta).

https://doi.org/10.1016/j.ecresq.2019.10.013
0885-2006/

Theory informed our basic premise that "good teaching is good teaching" across the many permutations in which it takes place. So far we see support for this. Remaining partly anchored in theory helped our work stay focused on interaction, and to build knowledge as well as tools. TTI/CLASS does not cover all teacher–child interactions or all elements of "quality." But we have learned that common, latent features of teacher–student interactions are resources for children and can be measured.

2. Interpreting effects of quality and measures

Another premise for the TTI and CLASS was the idea that features of "quality" should support children's development. Although hundreds of studies report significant relations between measures of quality, such as CLASS, and student outcomes (e.g., Mashburn et al., 2008; NICHD Early Child Care Research Network (ECCRN), 2002; Sabol, Hong, Pianta, & Burchinal, 2013), these effect sizes, when significant, are always modest (in the range of .05–.10) and in many instances are non-significant (Brunsek et al., 2017; Burchinal, Kainz, & Cai, 2011; Perlman et al., 2016). Evidence from causal designs that include random assignment of children to teachers show CLASS with significant, small causal effects of teacher–child interaction on learning (Carneiro, Cruz-Aguayo, & Schady, 2019). Reports of modest or no association(s) with child outcomes rightly prompt calls to develop new and improved measures of quality, as the field should. In considering new and improved assessments, obtained effect sizes should be interpreted in light of three realities: assessed child outcomes; multiple elements of quality; and constraints on education effects.

2.1. Outcomes

Nearly all studies of quality, including CLASS, use standardized assessments of outcomes distal to observed quality inputs. Such assessments (e.g., Woodcock & Johnson, 1989) capture generalized performance (e.g., achievement) and are not aligned to specific instructional experiences. Perhaps it is not surprising that assessments of quality would demonstrate modest effects on such generalized instruments. Interestingly, in the upper grades, when teachers align instruction to criterion-based standards assessments, validity coefficients for CLASS tend to be larger (Allen et al., 2013). It will be informative to evaluate effect sizes for quality measures, including CLASS, against a range of assessment types.

2.2. Multiple elements of quality

Evidence is increasing for a "package" of quality elements: teacher-student interactions, dosage of content, use of a targeted curriculum, and alignment of instruction to students' skills. In a recent investigation, this package was each independently and additively predicted children's learning, were uncorrelated, and yielded a larger effect size than each individually (Pianta et al., in press). Quality, even when defined by direct influences on development, is multi-faceted, and it is unlikely that any single quality assessment will measure all of them well. Further work on a package of quality elements is likely to advance much deeper understanding of classroom processes and their impacts.

2.3. Constraints on interpreting effect sizes

Interpreting effect sizes for quality should be informed by overall effects for "education." In studies of child development (e.g., NICHD ECCRN, 2002), "value-added" of education-related factors (Nye, Konstantopoulos, & Hedges, 2004), and causal impact of schooling (Carneiro et al., 2019), the *ceiling* effect size of education almost never exceeds .30 and is usually much lower. Genetics, family processes, and demographics account for much more variance while classrooms have the largest educational effects (Nye et al., 2004). When examining effect sizes of quality measures, we should perhaps value their magnitude relative to the education ceiling; maybe the denominator should change.

3. Measuring adult–child interaction: grain size

A first-order decision for any measure of quality is grain size; at what level of analysis (molar to molecular) will the tool focus? For CLASS, we faced whether to observe abstract "serve and volley" properties of interaction or in reduced form as discrete units. The *grain size* question reveals a tradeoff between reliability and validity. Theory informed our focus for CLASS at a molar level, the tradeoff being to define abstract properties so they could be observed and rated consistently. Typically, reliability for discrete units is easier to obtain (a smile is a smile), while molar codes rely more on observer judgement. For interactions, molar codes preserved meaning, and molecular grain size sacrificed meaning for agreement. In the TTI/CLASS taxonomy operational definitions across grain sizes (behavior, indicator, dimension, domain) present testable hypotheses about grain size tradeoffs against validity, reliability, efficiency, or theory. Our team is addressing these questions using TTI behavioral descriptors, master-coded exemplars, and scored tapes to examine ease of scoring, reliability across levels of analysis, and latent structure. We suspect results will inform efforts to build efficient, effective online observer training.

The tradeoff between reliability and validity and the problem of grain size are salient in large-scale applications. Questions arise around acceptable levels of error for measures used in high-stakes accountability. Should that error band differ when accountability is sanction-oriented (e.g., a teacher loses their job) or improvement-oriented (e.g., a teacher is assigned to training)? Another implication of grain size is the tendency of organizations (and people) to produce the units that are measured. For example, one of us worked with a school system using district-wide observations to evaluate teachers, the results having high-stakes consequences. The system counted teachers' use of open-ended questions, with three being the definition of a high score. In less than a year, every teacher used three open-ended questions. Because it is easy to display discrete behaviors, the evaluation became more of a compliance activity than a gauge of authentic teaching performance. Aligning the meaning of a score and its technical strengths and weaknesses, with its applied use, is a critical undertaking.

4. Measuring quality at scale

Early childhood education is used to collect data on classroom or program quality, witness the success and influence of the ECERS, with demand for quality metrics for tens of thousands of classrooms and reinforced by regulations such as in Head Start or Quality Rating and Improvement Systems. Experiences in two major studies provided opportunities for understanding and building operational infrastructure for use of observations like CLASS at scale: the Study of Early Child Care and Youth Development and the Multi-State Pre-K study. Experience in these studies gave RCP an appreciation for how working at scale influences measurement design as well as the need for infrastructure. Experience since clarified that working at scale is very different from even the largest sample research study.

For example, in these two studies the protocols called for classroom visits on a particular day and time, meaning we had to ensure that "quality" could be defined, operationalized, and assessed similarly across highly variable circumstances. Even though we

scheduled observations for "typical instructional times" it was stunning to see the variations in actual practice. Instructional time in literacy ranged across worksheet literacy drills, storybook reading, child-driven center time, and children wandering around. Variation is the rule in United States classrooms and applying a "common" denominator must capture shared properties. For CLASS, the common feature was teacher–student interaction, and we coded interactions at the molar level to support application across such varied instances.

Working at scale involves operations—training hundreds of observers, maintaining reliability, delivering data to users. We had published two papers using CLASS in the NCEDL (e.g., Mashburn et al., 2008) and within months were asked to scale—training hundreds of observers in response to new Head Start quality regulations. We were inventing infrastructure for scale when we had only used the instrument in a large study.

We have a better understanding now than we did a decade ago of the infrastructure enabling rigorous observations at scale, and there are over 30,000 observers who have been trained and certified as reliable on CLASS. This has required dedicated efforts to design, engineer, build, and sustain the technical and personnel resources to implement a standardized protocol with fidelity, which are efforts different from the program of research on CLASS. For professional development tools associated with CLASS, the issues and needs are similar. Opportunities to work at scale have forced us to develop entirely new and different systems and tools for practitioners. If I (RCP) conclude anything from this work it is that most researchers vastly underestimate the differences between the largest study and sustainably applying tools at scale, with school districts, states, regional agencies and other organizations that each have different profiles of interest and capacity.

5. A summary and the future

Our last 15 years of work on measuring quality reflect a dialectic between theory and application; we are interested in expanding a theory of teacher-student relationships and their value, and with intention to build and disseminate usable tools and knowledge. We believe it important for the field that this theory-application dialectic be reflected in all efforts to understand and improve early education and its impact. Looking ahead, we are intrigued by technology (natural language processing, artificial intelligence) that can make tools more efficient in terms of time and expense, and more effective. We hope to expand CLASS to capture other elements of quality efficiently and effectively, and to build system-level scaffolds for data collection, delivery, and workforce development that aid in the professionalization of the field and its stature as a national asset. Theoretically, as we work across cultures and countries, we are interested whether some properties of relationships are universal assets for development, and how culture-specific variations work. The agenda is exciting and the work is rewarding.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the Classroom Assessment Scoring System—Secondary. *School Psychology Review*, *42*(1), 76–98.

Brunsek, A., Perlman, M., Falenchuk, O., McMullen, E., Fletcher, B., & Shah, P. S. (2017). The relationship between the Early Childhood Environment Rating Scale and its revised form and child outcomes: A systematic review and meta-analysis. *PLoS One*, *12*(6), e0178512

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well are our measures of quality predicting to child outcomes: A meta-analysis and coordinated analysis of data from large scale studies of early childhood settings. In M. Zaslow, K. Tout, T. Halle, & I. Martinez-Beck (Eds.), *Next steps in the measurement of quality in early childhood settings*. Baltimore: Brookes Publishing.

Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2019). *Experimental estimates of education production functions: Sensitive periods and dynamic complementarity*. Institute for Fiscal Studies, University College London.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., . . . & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *Elementary School Journal*, *113*(4), 461–487. http://dx.doi.org/10.1086/669616

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257–1274. http://dx.doi.org/10.1111/cdev.12184

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.

NICHD Early Child Care Research Network. (2002). Child-care structure → process → outcome: Direct and indirect effects of child-care quality on young children's development. *Psychological Science*, *13*(3), 199–206.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257. http://dx.doi.org/10.3102/01623737026003237

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the classroom assessment scoring system) in early childhood education and care settings and child outcomes. *PLoS One*, *11*(12), e0167660.

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *Classroom assessment scoring system—PreK [CLASS]*. Baltimore: Brookes Publishing.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. http://dx.doi.org/10.3102/0013189X09332374

Pianta, R. C. Whittaker, J. E., Vitiello, V., Ruzek, E., Ansari, A., Hofkens, T., & DeCoster, J. (in press). Children's school readiness skills across the pre-K year: Associations with teacher-student interactions, teacher practices, and exposure to academic content. *Journal of Applied Developmental Psychology*.

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.

Sroufe, L. A. (1996). *Cambridge studies in social & emotional development. Emotional development: The organization of emotional life in the early years*. New York, NY: Cambridge University Press.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson psycho-educational battery—Revised*. Allen, TX: DLM Teaching Resources.