Paper Title Exploring Variance in the Pre-K CLassroom Assessment Scoring System (CLASS) Across Classroom Contexts

Author(s) Martha J. Buell, University of Delaware; Myae Han, University of Delaware; Henry May, University of Delaware; Carol Vukelich, University of Delaware

Session Title Measuring Early Childhood Classroom Quality

Session Type Paper

Presentation Date 4/30/2013

Presentation Location San Francisco, California

Descriptors Classroom Assessment, Early Childhood, Preschool

Methodology Quantitative

Unit SIG-Early Education and Child Development

Exploring Variance in the Pre-K Classroom Assessment Scoring System (CLASS) Across Classroom Context

By

Martha J. Buell

Henry May

Myae Han

Carol Vukelich

University of Delaware

April 30, 2013

Paper presented at the annual meeting of the American Educational Research Association

San Francisco, California

Introduction

First used in the National pre-K study (Pianta, Howes, Burchinal, Bryant, Clifford, Early, & Barbarin, 2005; Howes, Burchinal, Pianta, Bryant, Early, Clifford, & Barbarin, 2008), the CLASS™ was developed as a measure of the process quality of a classroom looking specifically at the quality of teachers' instructional practices. Unlike the Environmental Rating Scales (ECERS-R: Harms, Clifford, Cryer, 2005, ITERS-R: Harms, Cryer, Clifford, 2008, etc.) that are deemed to provide weight to the materials in a room, the CLASS™ is designed to evaluate the teacher's instructional strategies and skills, regardless of the wealth of the center. And while the dimension of teacher practice associated with responsivity to children remains, the CLASS is different from the Observational Record of Classroom Environment (OCRE:NICHD, 2002) in that the CLASS measures the teacher's practices regarding the classroom as a whole, rather than looking at practices associated with particular children. Because there is a critical need for the evaluative data that the CLASS ™ can supply, since its introduction in 2002 (Laparo, Pianta, Hamre & Stuhlman, 2002) the use of the CLASS ™ has grown, and in 2007 the CLASS ™ became an integral part of the Head Start monitoring system

> *"The Improving Head Start for School Readiness Act of 2007 requires that the Office of Head Start (OHS) include in the monitoring reviews of Head Start agencies a valid and reliable research-based observational instrument that assesses classroom quality, including the assessment of multiple dimensions of teacher-child interactions that are linked to positive child outcomes and later achievement. The conference report accompanying the Act suggests that OHS*

> *consider using existing research-based methods such as the Classroom Assessment Scoring System (CLASS) for this purpose."*

Initiated in 1965, Head Start is one of the last remaining weapons still being deployed in the United State's "War on Poverty". And since the 1969 Westinghouse Report (Westinghouse Learning Corporation and Ohio University, 1969), the efficacy of Head Start and its success in ending poverty by increasing the educational outcomes of children living in poverty has been debated. What is not in debate is that the quality of the Head Start programming does have an influence on the outcomes of the children attending the program. Therefore, the use of CLASS™ as an instrument to measure program quality will influence the kinds of programming that the youngest and poorest of our citizens receive.

But what do we really know about the psychometric properties of the CLASS™? As the utility of the CLASS™ switches from a research instrument to an accountability tool, we must continue to explore its reliability and validity in the many ways the field is beginning to use the tool . Although prior psychometric work on the CLASS™ was done largely in no-stakes circumstances (Hamre, Pianta, Mashburn & Downer, 2007), the results suggest good reliability and predictive validity. For example, the data on the use of the CLASS™, indicates that children who attended classrooms that score higher on the CLASS™ go on to make higher scores and greater gains in academic and social emotional assessments than do children who attended programs that rated relatively lower on the CLASS™ (Curby, LoCasale-Crouch, Konold, Pianta, Howes, Burchinal, Bryant, 2009).

In national research using the CLASS™ general classrooms were rated more highly on emotional supportiveness and least highly on instructional support (Pianta, Howes, Burchinal, Bryant, Clifford, Early, Barbarin, 2005). According to Justice, Mashburn, Hamre, & Pianta, (2008), even in classrooms where teachers were implementing a prescribed curriculum with fidelity, if the interactions were not scored highly on the CLASS™ instrument, the children did not do as well in measures of their literacy skills.

While the CLASS™ has been widely used for research on quality and child outcomes, a considerable amount of the extant research in peer-reviewed publications has used the CLASS™ in the context of lessons delivered in a large-group format. However the activities and experiences children have in preschool classrooms extend far beyond whole-group instruction, and include center time, small-group time, meal times and outside-play time. These are all important components of a high-quality program, in addition to large group/circle time and they are all important times for learning that could benefit from quality instructional support. However much less is known about how these components of the preschool daily schedule may influence CLASS™ scores. Therefore this study is designed to look at the variance in CLASS™ scores across three different instructional settings: large-group time, small-group time and center time. While the large group and in some cases the small group formats are similar to those used in the data published on the CLASS™, center time is less widely examined in the peer-reviewed research.

Theoretical Framework. We are grounding this work in a Generalizability Theory (G-theory) framework (Cronbach, Nageswari, & Gleser, 1963), using a multi-faceted G-theory model.

According to Webb, Shavelson and Haertel (2006), G-theory allows one to identify the sources of measurement error, disentangle them, and estimate the effect of the source of error in the resulting measurement. While the object of measurement (i.e., teachers in this case) is not considered a source of error, other features of a measurement situation that contribute to variability in scores are called facets. Thus, in this study we are considering the facets of *context* (i.e., large group, small group, center time), CLASS**™** *dimension, and time* (i.e. measurement occasion) as possible sources of variance or error in a CLASS™ score. In this way we can better understand how to maximize reliability of the CLASS**™** and recognize the possible threats to the validity of the CLASS™ instrument when it is used for the purposes of coaching or teacher professional development as separate from program or teacher evaluation. In this research study the influences of the context of data collection is of particular interest. For, if CLASS™ scores are influenced by the context in which they are gathered, this feature of the CLASS™ must be taken into account when administering the instrument.

Method

Participants. Thirty-one pre-school teachers involved in two related Early Reading First projects implemented in a mid-Atlantic state were the subjects of this investigation. All but one of the teachers was female. Sixteen of the teachers were Caucasian, eleven were African American and four were Latina.

The teachers were involved in an Early Reading First project, which used the CLASS™ as a tool to help coach the teachers on improving their classroom effectiveness. As such, the videos were collected on the teachers throughout the day in accordance with the format prescribed by the CLASS™ developers. The videos of the teachers were then were sent to an external lab where coders trained to maximize reliability on the CLASS™ coded the videos.

Instrument: Each video segment was scored according to the 7- point CLASS™ ratings for each of the 10 CLASS™ dimensions. At the time of this research the CLASS™ had ten dimensions: Positive climate, Negative climate, teacher sensitivity, regard for student perspective, behavior management, productivity, instructional learning format concept development, quality of feedback and language modeling. These 10 dimensions can be further reduced into three domains: Emotional Support, Classroom Organization, and Instructional Support. Table 1 depicts the three domains and the dimensions in each.

Table 1. *CLASS Dimensions and Domains*

Domain	**Emotional Support**	**Classroom Organization**	**Instructional Support**
Dimension	Positive Climate Negative Climate Teacher Sensitivity Regard for Student Perspectives	• Behavior Management • Productivity •Instructional Learning Formats	•Concept Development •Quality of Feedback •Language Modeling

Time.The data used for this analysis were collected over the course of two years. The first data collection point was in the fall or winter, the second in winter or spring, and the third occurring in the spring or summer. In some cases there were three samples collected, but in the case of center times, a fourth data collection point occurred the following fall.

In our project, we used the coded to tapes to assess our progress in assisting the teachers in creating language and literacy rich environments. Because classroom teaching occurs throughout the day, we collected the teaching samples across three classroom contexts, large group, center time and small group. In *large group* time teachers implemented several language and literacy strategies that we focused on throughout the ERF project, these included dialogic reading strategies, phonological awareness activities and alphabet recognition activities. During *center time* our focus with the teachers was on building target vocabulary words as well as having deep conversations through playful interactions in theme based activity centers, in which children interacted with learning centers such as block, dramatic play, library, math/science, art and the like. In addition our project focused on the benefits of dramatic play, and as such we tried to assist the teachers in supporting children's narrative imaginative paly throughout the day. In the *small group* time we had the teachers focus on various literacy skill building activities such as letter formation, recognition and sound, writing activities, phonological awareness activities and some play based vocabulary instruction. The literacy coaches used the videotapes collected on the teachers, as well as the scores that were given to the video samples in their coaching meetings designed to enhance the teachers' classroom skills and practice. We used the CLASS™ dimensions and structure as a guide to additional skills and teaching development, with the hope that teachers' scores on the CLASS™ instrument would increase. However, for this analysis the individual teachers growth was not assessed, rather we wanted to study the ways that the CLASS™ scores were influenced by the context in which they were collected, and if there was a difference in the CLASS™ domain depending on context.

Analysis. Table 2 contains the mean scores and standard deviations by the context in which the data was collected. We present both the overall mean and standard deviation by context, and then subdivide the scores by domain and context.

Table 2. *Class Dimension/Domain by Classroom Context*

Context	CLASS™ Dimension Domain	Mean	Standard Deviation
Large group n=65	Overall	4.35	1.12
	emotional	4.98	.825
	instructional	3.11	1.10
Center time n=74	Overall	4.64	1.15
	emotional	5.48	.95
	instructional	2.93	1.16
Small group n=29	Overall	4.80	1.24
	emotional	4.96	.95
	instructional	2.62	1.12

Next, we estimated a multifaceted G-Theory model using basic hierarchical linear modeling (HLM) techniques (Raudenbush & Bryk, 2002). Individual item scores were modeled as a function of teacher, CLASS™ dimension, context, and occasion effects. The main effects for contexts and dimensions were included as fixed effects, since the three contexts and ten CLASS™ dimensions represent the full scope of measurement (i.e., they are not randomly sampled from a larger population of contexts and dimensions) and the average rating across all teachers may differ by context and dimension. Teacher and occasion effects, along with teacher by context and teacher by dimension interactions were included as random effects, with variance component estimates reflecting the relative amounts of error variance associated with each of these facets. This variance components model was estimated using restricted maximum likelihood (REML) via PROC MIXED in SAS 9.3.

Results

Estimates for the error variance components are shown in the table below. All variance components were statistically significant at the 99% level or better. More importantly, the error variance component for *context* was larger than that for *dimension*, and nearly as large as that for measurement *occasion*. This suggests that it is not sufficient to simply collect CLASS™ data on multiple occasions. We must also pay close attention to the context for each data collection occasion and ensure that data for each teacher is collected across multiple contexts on multiple occasions. Table 3 provides the estimates of the error variance, standard error and total error for CLASS scores

Table 3. *Error variance by dimension, context and time*

	Estimate	Standard Error	Percent of Total Error
Dimension	0.1025	0.0221	9.2%
Context	0.1545	0.0407	13.9%
Time	0.2215	0.0554	20.0%
Residual Error (i.e., item-level)	0.6308	0.0262	56.9%

Discussion

Head Start strives to provide the best possible learning environments for pre-school children living in poverty. And while Head Start is perhaps the oldest, adding a Pre-K year as a means of supporting the success of low-income children is growing as a state and federal investment (www.preknow.org). As public dollars increase, the need for accountability provided in part by monitor the quality of preschool programming also increases. In recent years the CLASS™ has become widely used in order to both evaluate the quality of classrooms and programs, and as a tool for improving classroom quality. In particular the CLASS™ has been identified as a means to improve teacher practice. For example, the My Teaching Partner (Mashburn, Downer, Hamre, Justice, & Pianta, 2010; Pianta, Mashburn, Downer, Hamre, & Justice, 2008) program that uses videotapes submitted by teachers in order to provide feedback on teaching practices based on how well the teachers performed according to the dimensions of the CLASS™ in order to improve practice. However the current data suggest that a more nuanced approach must be taken both towards evaluation and quality improvement when using the CLASS™. Our analysis of the CLASS™ data indicate that the context within which activities are being done at the time of the sampling adds considerable variance to a teacher's CLASS™ score. This feature of the CLASS™ measure should be further studied in a larger sample of teachers, but with similar controls for context. Of course our sample is small and there is a need for further study, but these data indicate that perhaps the nature of activities themselves might affect the scores. This finding is critical from both an evaluation as well a quality enhancement perspective. From the evaluation perspective, ensuring that there are proportionally equivalent amounts of time spent in each context in order to evaluate teacher abilities would be advisable. In these data, small group was the context with the overall highest mean score, and the large group circle time had the lowest. While this finding should lead to caution and perhaps guidance in how to collect samples of CLASS™ data in order to evaluate the overall quality of a program, the domain by context was equally compelling when one considers there use in quality enhancement efforts. For example, when the scores were further deconstructed by Domain, the emotional support domain was highest in center times and lowest in large group. Conversely, the instructional strategies domain was highest in Small group and lowest in center time. The very nature of the activities that are appropriate to happen in these various areas – small group being more didactic, skill focused and teacher driven and center time being more child driven are likely explanation for the differences. However, ensuring that as teachers are offered support for their practice ensuring that samples are taken across contexts really enhances and clarifies not only global growth, but also targeted differences. It is quite likely that some teachers may be less skilled in providing instruction in a playful manner during a center time; likewise there are teachers that may struggle with supporting children's social and emotional needs when delivering targeted instruction. Therefore, knowing the teachers' strengths and areas of needed growth, across different contexts allowed our project to be much more precise in providing teachers with feedback and coaching.

Another important component to consider is the impact of timing in the collection of data. While we would like to say that the more time the teachers spent with our program the higher their scores became, that is not a significant finding. Based on these data the scores fluctuated by day, but not in a consistent way. Please see Table 4 for a comparison of CLASS™ scores by context over time.

Table 4. *CLASS scores by context across time*

	Time 1	Time 2	Time 3
Large group	4.213	4.47	4.40
Center time	4.40	5.38	5.19

There are many possible explanations for why the scores fluctuate overtime. Possible sources of variance could be the children in the classroom at the time of the measure, the teaching team, or the time of year. Further research on the aspects of the classroom environment aside from the teacher practice need to be further studied as these features may also be critical to control for as the CLASS™ is used to assess early childhood classrooms and programs.

References

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., et al. (2009). The Relations of Observed Pre-K Classroom Quality Profiles to Children's Achievement and Social Competence. *Early Education & Development*, 20(2), 346–372.

Hamre, B., Pianta, R., Mashburn, A., & Downer, J., (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. Early childhood and elementary classrooms. Downloaded on March 27, 2013 from http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf.

Harms, T., Clifford, R.M., & Cryer, D. (2005). *Early Childhood Environmental Rating Scale: Revised Edition.* New York: Teachers College Press.

Harms, T., Clifford, R.M., & Cryer, D. (2008). *Infant Toddler Environmental Rating Scale: Revised Edition.* New York: Teachers College Press.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. Early Childhood Research Quarterly, 23(1), 27–50.

Justice, L. M., Mashburn, A. J., Hamre, B. K., & Pianta, R. C. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly*, 23(1),51–68.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre- kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science*, 9(3), 144–159.

Pianta, R.C., La Paro, K., & Hamre, B. K. (2008) *Classroom Assessment Scoring System*. Baltimore: Paul H. Brookes.

Pianta, R.C., Mashburn, A. J., Downer, J. T., Hamre, B. K. & Justice, L. (2007). Effects of web- mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 431-451.*

Pianta, R. C, Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431–451.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the Prekindergarten Year. *The Elementary School Journal*, 104(5), 409.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., Burchinal, M., Early, D., & Howes, C. (2008). Pre-k program standards and children's development of academic, language, and social skills. *Child Development*, *79*, 732-749.

Mashburn, A., Downer, J., Hamre, B., Justice, L., & Pianta, R. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, *14*(4), 179–196.

NICHD Early Child Care Research Network. (2002). Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Education Research Journal*, 39, 133 – 164.

Webb, N., Shavelson, R., Haertel, E. (2006). Reliability coefficients and Generalizabilty theory. *Handbook of statistics*, vol. 26, 81-124, Elsevier.

Whitaker, S., Kinzie, M., Kraft-Sayre, M. E., Mashburn, A., & Pianta, R. C. (2006). Use and Evaluation of Webbased Professional Development Services Across Participant Levels of Support. *Early Childhood Education Journal*, 34(6), 379-386.

Westinghouse Learning Corporation and Ohio University. (1969). The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development (Vols. 1 and 2, Report to the Office of Economic Opportunity). Athens: Authors.