Journal of Psychoeducational Assessment
2016, Vol. 34(8) 765–781
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0734282915627115
jpa.sagepub.com
SAGE

Live Versus Video Observations: Comparing the Reliability and Validity of Two Methods of Assessing Classroom Quality

Timothy W. Curby[1], Price Johnson[2], Andrew J. Mashburn[2], and Lydia Carlis[3]

Abstract

When conducting classroom observations, researchers are often confronted with the decision of whether to conduct observations live or by using pre-recorded video. The present study focuses on comparing and contrasting observations of live and video administrations of the Classroom Assessment Scoring System–PreK (CLASS-PreK). Associations between versions, mean differences, reliability, and predictive validity were examined. Results generally indicated high correlations between versions. Video codes were slightly lower on average than live codes. Reliability was generally acceptable in terms of Cronbach's alpha, but multigroup confirmatory factor models suggested some differences between observation types. Finally, CLASS scores based on each observation type indicated some predictive validity of children's academic achievement, but no observation type was uniformly better. The discussion focuses on why the codes might differ and the implications of those differences.

Keywords

Preschool observations, measurement, validity, reliability

Classroom observations are increasingly being used in education research, teacher professional development programs, and teacher performance evaluation systems to capture teacher and program quality (Joe, Tocci, Holtzman, & Williams, 2013). Districts and schools may use these standardized protocols as a way to rate teacher performance, and these ratings are sometimes associated with high-stakes decisions (e.g., raises, bonuses, or losing jobs). Researchers remain interested in these measures because they assess features of classrooms that are proximal to the developing child and that are predictive of a variety of student outcomes (Mashburn et al., 2008). Because of the high stakes, there is growing interest in understanding how observational instruments may function under different conditions. One common decision in using these instruments is whether to conduct the observations live in classrooms or via a video from the classroom. Live

[1]George Mason University, Fairfax, VA, USA
[2]Portland State University, OR, USA
[3]AppleTree Institute for Education Innovation, Washington, DC, USA

Corresponding Author:
Timothy W. Curby, George Mason University, 4400 University Dr., MS 3F5, Fairfax, VA 22030, USA.
Email: tcurby@gmu.edu

coding has the main advantage of the observer being able to take into account everything that is happening in the classroom, and not just what is visible and audible on a video. Video coding has the possible advantages of being done asynchronously, being done more cheaply (e.g., teachers capture videos themselves), and potentially being coded multiple times. Thus, researchers may have good reason to choose either method. Although studies have found evidence that scores from each method have predictive validity for child outcomes (e.g., Curby et al., 2009; Mashburn et al., 2008), research has yet to identify if there are differences between live versus video observations in predicting outcomes. In other words, might the results of a given study be an artifact of what observational method was used? The present study focuses on comparing and contrasting observations of live and video administrations of the Classroom Assessment Scoring System–PreK (CLASS-PreK; Pianta, La Paro, & Hamre, 2008) and how scores from each method may differentially relate to child academic achievement.

Observations of Classroom Quality

There is an increasing emphasis on classroom observation as a metric of teacher and program quality. Recent education policies, such as Race to the Top, explicitly encourage the use of teaching observations (along with other indicators) to evaluate teacher performance (U.S. Department of Education, 2009). Ratings from these observations may be used in a variety of capacities. Lower stakes decisions often include identifying areas of challenge for teachers for professional development or research purposes (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Mashburn, Downer, Hamre, Justice, & Pianta, 2010). High-stakes decisions may include deciding which teachers should be fired, teacher bonuses, reimbursement rates an early childhood center receives for each child, as well as the number of stars an early childhood center receives on a quality rating and improvement system (QRIS; Tout, Zaslow, Halle, & Forry, 2009).

There are a variety of factors that explain why observational instruments are increasingly being used in classrooms. First, standardized instruments can differentiate between schools, classrooms, and teachers. Regardless of the instrument, classroom observations have value because outside observers can give ratings that allow for comparisons to a set standard. Thus, comparisons can be made across classrooms, schools, districts, and so on—a property that is of great value to researchers, policy makers, and parents (in the case of QRIS ratings). In other words, scores from psychometrically validated observational measures should reflect real differences between classrooms.

Second, classroom observations allow for evaluations of classrooms that are not reliant on student test score data. The other ascendant approach in evaluating teacher performance is to use value-added models using student test score data (National Research Council and National Academy of Education, 2010). However, as conceptually appealing as the proposition of using student data is, the challenges associated with using those data are substantial (American Statistical Association, 2014). For example, the changes in student outcomes attributable to teachers from value-added models are not very reliable with only a few years worth of data. Conversely, evaluations based on classroom observations do not directly take into account student test score data, but rather look at how teachers interact with children, regardless of the content being taught.

Third, research using classroom observations has linked observed classroom quality to student social and academic outcomes (Bell et al., 2012; Bill and Melinda Gates Foundation, 2012; Mashburn et al., 2008). Although the focus of the present study is the CLASS-PreK instrument, a variety of observational instruments have been developed and used in preschool classrooms to predict child outcomes. For example, the Early Childhood Environment Rating Scale–Revised (Harms, Clifford, & Cryer, 1998) and the Arnett Caregiver Interaction Scale (Arnett, 1986) are two other global observation systems that are predictive of children's outcomes.

However, knowing that an instrument is predictive of outcomes is not enough. Because the classroom ratings are being used in high-stakes decisions, procedural differences between different administration types may be important to consider. One procedural decision has to do with

using live or video observations. Live coding involves sending a rater to a classroom to conduct the observations. Usually, the rater sits unobtrusively in a part of the room that provides both clear visibility of the teacher and the ability to hear his or her interactions with students. By being in the classroom, the observer has the flexibility to look around as well as the ability to move if she or he cannot hear the teacher. Thus, live observations may offer greater validity in the sense that they can capture everything that is happening in the classroom, and not just what is visible on a video.

Video coding involves setting up a video camera to capture the classroom. Sometimes the video camera has a person who keeps the camera focused on the teacher. Sometimes, the camera is set up (e.g., by the teacher) and has a set focal point. Either way, the video files that are produced have several advantages. First, there is also the possibility of the videos being coded more cheaply. Video files can be queued and rated over longer periods of time, which has the possibility of reducing training costs. Second, classroom videos can be uploaded to a site (Downer, Kraft-Sayre, & Pianta, 2009) and transmitted long distances for minimal costs, whereas sending personnel long distances for live coding can be costly. Third, if the teacher is wearing a microphone, the video may allow for the coders to hear speech that would be inaudible to an observer in the room. Fourth, either initially or later, videos have the possibility of being coded multiple times. In such cases, scores can be averaged across raters and reliability can be improved. Thus, researchers may have good reason to choose either method.

The present study uses the CLASS-PreK (Pianta et al., 2008), a widely used observational measure of the quality of teachers' interactions with children. There are more than 30,000 certified CLASS observers across grade levels worldwide (Teachstone, 2015). In fact, the CLASS has ballooned in popularity in part because of Head Start programs—a federally funded preschool program for at risk children—explicitly using the CLASS to measure program quality (Office of Head Start, 2014).

The CLASS-PreK manual identifies 10 dimensions of teacher interactions that are each rated on a 1 to 7 scale after observing a classroom for 20 minutes. Each of the 10 measured dimensions is further specified by four to five indicators, which articulate subcomponents that constitute that dimension. For example, the Positive Climate dimension of the CLASS comprises four indicators: relationships, positive affect, positive communication, and respect. All CLASS-PreK domains, dimensions, and indicators are presented in Table 1. Typically, researchers score at the dimension level (e.g., Positive Climate), which is commensurate with the CLASS training and manual. However, for those interested in professional development, a dimension can prove to be too broad of a construct for specific feedback, and thus, school personnel may prefer to assess at the indicator level (e.g., relationships). The present study uses ratings of these indicators. For comparison purposes, indicators can be aggregated into dimensions, and dimensions can be aggregated into domains.

Users of the CLASS, as well as other observational instruments, are confronted with the choice of whether to do the observations live or to base the ratings on videos of the classroom. The CLASS manual provides instructions for either form of observation. Instructions for video coding caution raters to code only what is visible on the video and that no inferences can be made about what is happening off screen. For example, if there is audible crying, but the source cannot be seen, that crying should not be taken into account in any ratings. Anecdotally, raters often prefer to conduct the observations live because context and meaning are easier to deduce with a wider field of vision than most camera setups and the ability to simply turn their heads to look. Thus, consistent with what many raters believe, live observations may potentially provide more valid ratings. However, video observations provide an easier mechanism to multiply-code segments and relieve some of the practical burdens of ratings, such as being able to more easily distribute the timing of the ratings (without necessarily distributing the time of the observation itself) and being able to rely on fewer coders. Research using a version of the CLASS for secondary classrooms (Pianta, Hamre, Haynes, Mintz, & La Paro, 2007) found that dimension (e.g., Positive Climate) and domain scores (e.g., Emotional Support) were generally higher for live codes than for video codes (Casabianca et al., 2013). It is not known whether some findings using the CLASS are reliant on whether classrooms were live- or video-coded.

Table 1. CLASS-PreK Domains, Dimensions, and Indicators.

Emotional support	Classroom organization	Instructional support
Positive climate	**Behavior management**	**Concept development**
Relationships	*Clear behavior expectations*	*Analysis and reasoning*
Positive affect	*Proactive*	*Creating*
Positive communication	*Redirection of misbehavior*	*Integration*
Respect	*Student behavior*	*Connections to the real world*
Negative climate	**Productivity**	**Quality of feedback**
Negative affect	*Maximizing learning time*	*Scaffolding*
Punitive control	*Routines*	*Feedback loops*
Sarcasm/disrespect	*Transitions*	*Promoting thought processes*
Severe negativity	*Preparation*	*Providing information*
		Encouragement and affirmation
Teacher sensitivity	**Instructional learning formats**	**Language modeling**
Awareness	*Effective facilitation*	*Frequent conversation*
Responsiveness	*Variety of modalities and materials*	*Open-ended questions*
Addresses problems	*Student interest*	*Repetition and extension*
Student comfort	*Clarity of learning objectives*	*Self- and parallel talk*
		Advanced language
Regard for student perspectives		
Flexibility and student focus		
Support for autonomy and leadership		
Student expression		
Restriction of movement		

Note. Emotional support, classroom organization, and instructional support are the domains. Boldface text indicates dimension. Italicized text indicates indicator. CLASS = Classroom Assessment Scoring System.

The purpose of this study is to explore video and live coding of an authentic assessment of preschool classrooms with the following research questions:

Research Question 1: To what degree are live and video ratings of the same classroom time related to one another?
Research Question 2: Are there mean differences in live and video ratings?
Research Question 3: Does internal consistency reliability vary across live and video assessments?
Research Question 4: Is predictive validity similar across live and video methods in predicting child academic outcomes?

Method

Participants

Classrooms consisted of lead teachers, resident teachers, and teaching assistants using the *Every Child Ready* curriculum developed by AppleTree Institute for Education Innovation in Washington, D.C. In total, observational data were available from 51 classrooms in 16 schools at five local education agencies. There were 95 teachers, resident teachers, and teaching assistants represented across participating classrooms, of whom 92.6% were female and 7.4% were male.

The majority of teaching staff had attained a bachelor's degree or higher (69.5%), and all classrooms were led by teachers with at least a bachelor's degree. On average, teaching staff reported 4.58 years of experience in the early care and education field.

From the 51 classrooms, there were 1,225 live-coded observations conducted by 22 raters. However, because raters were paired for the purposes of this study, and video ratings necessarily occurred after live observations, there were missing data for the video observations because three live observers did not continue on to be video coders. This reduced our sample observations to 769 live-video pairs of ratings.

The sample of students consisted of 593 children, 50.9% of whom were female ($n = 291$). In terms of ethnicity, 83.1% of students were African American ($n = 493$), 14% were Caucasian ($n = 83$), 1.7% were Asian ($n = 10$), 0.5% were Native Hawaiian/Pacific Islander ($n = 3$), and 0.7% were American Indian ($n = 4$). In addition, 3% of the students identified as Hispanic ($n = 18$), though this was not mutually exclusive with race. Nine percent of students were English language learners ($n = 60$), and 76.5% qualified for free or reduced price lunch ($n = 466$).

Measures

Classroom quality. The CLASS-PreK (Pianta et al., 2008) was used as a measure of the quality of interactions teachers have with children. The CLASS involves observing classrooms for approximately 20 min and then providing ratings on a 1 *low* to 7 *high* scale. The CLASS is organized such that there are three domains: Emotional Support, Classroom Organization, and Instructional Support. Emotional Support is made up of four dimensions: Positive Climate, Negative Climate (reversed), Teacher Sensitivity, and Regard for Student perspectives. Classroom Organization is made up of three dimensions: Behavior Management, Productivity, and Instructional Learning Formats. Instructional Support is made up of three dimensions: Concept Development, Quality of Feedback, and Language Modeling. Each of the 10 dimensions is made up of several indicators, which are what were scored by the raters in the present study (see Table 1).

Child outcomes. Several child outcomes were assessed at the beginning and end of the school year including the Test of Early Math Ability (TEMA-3; Ginsburg & Baroody, 2003), the Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 2007), and the Test of Early Preschool Literacy (TOPEL; Lonigan, Wagner, Torgeson, & Rashotte, 2007).

The TEMA-3 (Ginsburg & Baroody, 2003) is a norm-referenced measure that assesses children's mathematical abilities and conceptual understanding. The tool consists of 72 items that assess knowledge of concepts such as numbers, comparisons, addition, subtraction, multiplication, and division (Molfese et al., 2012). Test–retest reliability over a 2-week period has been reported as .82 (Molfese et al., 2012) while internal consistency reliability coefficients were found to be above .92 (Ginsburg & Baroody, 2003). Children's responses were scored on individual record forms, and then raw scores were transformed into math ability scores.

The PPVT-4 (Dunn & Dunn, 2007) is a measure of receptive vocabulary for standard American English. Children were shown four pictures and then asked to point to the one that best represents the word spoken by the assessor. Standard scores were then created by comparing participants' scores with a normative sample of children in the same 6-month age range taken from a larger sample of 3,450 individuals ranging from 2 years 6 months to 81 years of age (Dunn & Dunn, 2013). The PPVT-4 exhibits high test–retest reliability ($r = .92$-$.96$), split half reliability ($r = .94$-$.95$), and alternate-form reliability ($r = .87$-$.93$; Dunn & Dunn, 2013).

The TOPEL (Lonigan et al., 2007) is a measure of young children's emergent literacy that is comprised of three subscales: Print Knowledge, Definitional Vocabulary, and Phonological Awareness. Print Knowledge assesses children's alphabet knowledge (letter naming) and written language conventions and form (word identification, association of letter sounds with written

form, etc.). Definitional Vocabulary assesses children's single-word oral vocabulary and definitional vocabulary (the child is shown a picture and asked to describe its important features). Phonological Awareness measures children's word elision (the child is asked to speak after removing specific sounds from words) and blending abilities (the child listens to two separate sounds and is asked to combine them). The TOPEL has shown convergent reliability with similar measures (r = .59-.77) and is predictive of reading skills in both kindergarten and first grade (Wilson & Lonigan, 2009). Two-week test–retest reliability for the TOPEL has been found to range from .81 to .89. Internal consistency reliability for the three subsets ranges from .86 to .96 (Wilson & Lonigan, 2009).

Child demographics. Schools provided information on student gender, free and reduced price lunch status, minority status, and English language learner status.

Procedure

All classrooms were using the same *Every Child Ready* curriculum developed by AppleTree Institute for Education Innovation in Washington, D.C., with funding from a federal Investing in Innovation grant. The schools in this study use the CLASS observations for teacher evaluation and to inform professional development provided for teachers. Participating schools also use the CLASS as part of their accountability plans for the D.C. Public Charter School Board.

Raters were trained in the CLASS-PreK measure during a 2-day training led by a certified CLASS trainer. After training, raters had to rate five 20-min videos. Typically, raters provide ratings on each of the 10 dimensions. To be able to provide teachers with more specific professional development, a notable change to typical observation procedures was that AppleTree revised the coding of observations so that ratings were given for each of the indicators, instead of the dimensions. Because raters were going to have to provide ratings at the indicator level, but the videos are master-coded at the dimension level, raters' indicator codes were aggregated to the dimension level and then compared with the master codes for reliability certification purposes. As with any CLASS training, these dimension aggregates had to agree within plus or minus one scale point on 80% of dimension codes on five videos to conduct observations. Videos were also locally master-coded at the indicator level to provide raters with feedback at the indicator level during training.

Live raters brought a video camera with them to the observation. Each teacher was observed using the CLASS-PreK measure, on 2 days at 3 times throughout a single year. Observations took place on different days of the week and during different times of the day. While the live coding was taking place, a video camera was used by the rater to capture the classroom interactions. These videos were coded approximately 2 to 3 weeks later. Raters also kept a running script of what was taking place in the classroom. The coding of these videos mirrored the live coding's general structure. One member of a rating pair was assigned to live code Day 1 and the second member was assigned to live code Day 2. This pair of coders swapped the observations such that whoever live-coded a given teacher's Day 1 would video code Day 2, and whoever live-coded that teacher's Day 2 would video code Day 1. This procedure removed a notable confound related to the fact that rater variance will be pooled with classroom variance (thus rater effects are balanced across live and video observations) and allows for the most direct comparison between methods of coding.

External, independent contractors with a minimum of a bachelor's degree were hired and trained by the external evaluator and conducted beginning and end-of-year standardized assessments.

Gender, free and reduced price lunch status, minority status, and English language learner status were determined from school records.

Results

To What Degree Are Live and Video Ratings of the Same Classroom Time Related to One Another?

The analyses were used to study similarities and differences between the methods of coding. To answer our first research question, about the degree live and video ratings are related to one another for the same classroom time, paired correlations were computed at the indicator, dimension, and domain levels. No correction was made to the *p* values for the many correlations, as the focus of these analyses was descriptive in nature and thus focused on the magnitude of association, not the statistical significance. As shown in Table 2, correlations were generally small to moderate in size. Smaller correlations were evident in the indicators and the largest correlations were among the domains, with the dimensions generally in the middle. This suggests that as items were pooled and error was distributed across more and more items, reliability increased and the correlations were less influenced by small differences in the ratings. Results indicated that all indicators, dimensions, and domains are significantly correlated across live and video observations (Table 2), except for one indicator of Negative Climate (Severe Negativity) and one indicator of Concept Development (Connections to the Real World). The lack of correlation across live and video ratings of Severe Negativity is likely due to the very small degree of variance in those ratings. Given that Connections to the Real World had substantial variance in both methods, the lack of a significant correlation may suggest that this indicator is not being reliably coded in one or both methods.

Are There Mean Differences in Live and Video Ratings?

Notably, the correlations do not reveal whether there are mean differences in ratings, which was our second research question. It is possible for scores to be highly correlated (i.e., when live ratings are higher, video ratings are higher, too), but still be different (such as if one method provides lower ratings overall). Thus, *t* tests were also conducted to test for differences in the indicators, dimensions, and domain scores. Although the number of tests increases the likelihood of making a Type-I error, we descriptively wanted to give readers a sense for what is actually different versus what appears different, and thus we present the results without an error correction. The means from both live and video ratings would be considered typical of preschool classrooms, but significant differences emerged between the video-coded means and the live-coded means. In the Emotional Support domain, the Negative Climate dimension, as well as three of the four Negative Climate indicators, evidenced significant differences between live and video ratings (Table 2), with video scores being slightly lower (i.e., better) than live scores. In the Classroom Organization domain, only Student Behavior (an indicator of Behavior Management) was significantly different, and in this case, the video mean was higher than the live mean. In the Instructional Support domain, two indicators of Concept Development—Integration and Connections to the Real World—were significantly lower in the video coding. In addition, Language Modeling and two of its five indicators (Repetition and Extension, Self- and Parallel Talk) were significantly lower in the video coding.

Does Internal Consistency Reliability Vary Across Live and Video Assessments?

Our third research question asked whether reliability ratings varied across live and video assessments. This was particularly important given that indicators were scored within each dimension, and thus no other reliability ratings in other studies are available. This examination of reliability was done in two ways. The first was to compute Cronbach's alphas for the indicators in each

Table 2. Descriptive Statistics, Correlations, and *T* Tests for Live and Video CLASS Codes.

	Live			Video		Correlation	*t* test
	N	*M*	*SD*	*M*	*SD*	*r*	*t*
Emotional support	769	5.70	0.81	5.76	0.74	.43***	−1.95
Positive climate	768	5.74	1.10	5.80	1.00	.37***	−1.45
Relationships	768	5.76	1.21	5.80	1.07	.24***	−0.85
Positive affect	768	5.53	1.37	5.54	1.30	.35***	−0.17
Positive communication	768	5.67	1.34	5.80	1.14	.29***	−2.39
Respect	768	6.00	1.13	6.06	0.99	.30***	−1.46
Negative climate	769	1.28	0.52	1.17	0.32	.29***	6.06***
Negative affect	769	1.52	0.87	1.36	0.66	.33***	5.06***
Punitive control	769	1.38	0.79	1.22	0.55	.25***	5.19***
Sarcasm/disrespect	769	1.11	0.47	1.07	0.29	.17***	2.23
Severe negativity	768	1.11	0.59	1.03	0.17	−.03	3.68***
Teacher sensitivity	768	5.48	1.10	5.51	1.08	.30***	−0.58
Awareness	768	5.29	1.25	5.36	1.18	.22***	−1.26
Responsiveness	768	5.29	1.29	5.35	1.22	.28***	−1.13
Addresses problems	767	5.37	1.32	5.43	1.23	.28***	−1.08
Student comfort	767	5.97	1.15	5.89	1.11	.24***	1.65
Regard for student perspectives	767	4.50	1.34	4.56	1.24	.37***	−1.07
Flexability and student focus	766	4.60	1.62	4.49	1.60	.31***	1.70
Support for autonomy and leadership	767	4.57	1.65	4.53	1.62	.24***	0.63
Student expression	766	5.13	1.51	5.22	1.49	.26***	−1.31
Restriction of movement	767	5.18	1.65	5.33	1.41	.24***	−2.27
Classroom organization	769	5.37	0.99	5.40	0.97	.45***	−1.01
Behavior management	769	5.37	1.17	5.40	1.08	.36***	−0.77
Clear behavioral expectations	769	5.64	1.36	5.47	1.26	.19***	2.80
Proactive	769	5.07	1.45	5.12	1.27	.16***	−0.80
Redirection of misbehavior	769	5.21	1.44	5.30	1.34	.37***	−1.47
Student behavior	768	5.54	1.32	5.71	1.19	.39***	−3.52***
Productivity	768	5.46	1.16	5.53	1.07	.31***	−1.54
Maximizing learning time	765	5.54	1.39	5.56	1.35	.23***	−0.21
Routines	768	5.91	1.18	5.93	1.15	.31***	−0.40
Transitions	665	5.51	1.53	5.65	1.33	.21***	−1.99
Preparation	763	6.25	1.13	6.17	1.05	.29***	1.71
Instructional learning formats	768	5.06	1.21	5.10	1.24	.39***	−0.73
Effective facilitation	766	5.11	1.45	5.20	1.44	.39***	−1.66
Variety of modalities and materials	761	4.76	1.70	4.95	1.66	.28***	−2.60
Student interest	768	5.53	1.27	5.49	1.30	.31***	0.74
Clarity of learning objectives	760	4.83	1.80	4.76	1.81	.22***	0.93
Instructional support	767	3.45	1.16	3.36	1.04	.38***	2.02
Concept development	763	2.94	1.31	2.75	1.27	.16***	3.19
Analysis and reasoning	763	2.67	1.57	2.64	1.51	.21***	0.51
Creating	760	2.91	1.68	2.88	1.65	.22***	0.37
Integration	762	3.37	1.79	3.01	1.73	.12***	4.29***
Connections to the real world	762	2.90	1.75	2.54	1.59	.09	4.41***
Quality of feedback	766	3.32	1.32	3.46	1.22	.34***	−2.62
Scaffolding	761	4.16	1.70	4.23	1.59	.24***	−1.02
Feedback loops	765	3.35	1.71	3.39	1.63	.16***	−0.52

(continued)

Table 2. (continued)

	Live			Video		Correlation	*t* test
	N	*M*	*SD*	*M*	*SD*	*r*	*t*
Prompting thought processes	765	2.73	1.66	2.83	1.65	.17***	–1.32
Providing information	765	3.81	1.57	3.93	1.59	.29***	–1.81
Encouragement and affirmation	758	4.59	1.62	4.63	1.57	.28***	–0.55
Language modeling	765	3.29	1.28	3.13	1.13	.39***	3.26***
Frequent conversation	765	4.58	1.70	4.62	1.56	.25***	–0.51
Open-ended questions	763	3.13	1.75	3.09	1.68	.38***	0.63
Repetition and extension	763	3.91	1.58	3.61	1.53	.30***	4.43***
Self- and parallel talk	763	3.41	1.73	3.15	1.66	.22***	3.30***
Advanced language	759	3.32	1.78	3.13	1.66	.35***	2.62

Note. Emotional support, classroom organization, and instructional support are the domains. Boldface text indicates dimension. Italicized text indicates indicator.
*$p < .05$. **$p < .01$. ***$p < .001$.

dimension separately for live and video methods. Alphas are presented for the dimensions in Table 3. All alphas were above .70, except for the video coding of Negative Climate. As with the correlations, this is likely due to the especially small amount of variance in Severe Negativity and Sarcasm/Disrespect, particularly in video ratings of these indicators (which may not have the negative event within the frame of the picture). No clear patterns emerged with higher or lower alphas for one method or another.

Confirmatory factor analyses (CFAs) were also conducted to examine the reliability of the two methods. Nesting of observations within classrooms was accounted for in MPlus using TYPE = COMPLEX. Each model consisted of all the indicators within a dimension and the dimensions within a given domain. Unconstrained multigroup CFAs computed factor loadings separately for the live/video coding methods and are reported in Table 3. Then constraints were added to test to see if the factor loadings varied significantly from one method to another. We tested to determine whether constrained models fit significantly worse than the unconstrained model through nested model comparisons. These nested model comparisons provided a chi-square change value with its corresponding degrees of freedom change for each constrained–unconstrained model pair. If the constrained model fit significantly worse than the unconstrained model according to a chi-square difference test, then we determined that there was a significant difference between the factor loadings (Vandenberg & Lance, 2000). As a way to work through the potentially large number of comparisons, constraints were first done at the domain level (e.g., all factor loadings of the indicators within Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student Perspectives were constrained to be equal across video and live CFAs), and then constraints were conducted by each dimension (e.g., Positive Climate) to identify the areas where the dimensions varied from one another.

At the domain level, Emotional Support, Classroom Organization, and Instructional Support all showed significant differences between the live and video models (See the appendix). Post hoc analyses were conducted to investigate which dimensions were responsible for the significant differences in the domains. When a significant difference was found, this suggests that the factor loadings in the live and video models were significantly different. In terms of Emotional Support, Negative Climate and Regard for Student Perspectives both showed significant differences (Table 3). Classroom Organization showed differences in Instructional Learning Formats (Table 3). Instructional Support showed differences in both Quality of Feedback and Language Modeling (Table 3).

Table 3. Standardized Factor Loadings From Multigroup Confirmatory Factor Analyses.

	Live	Video	*p*
Emotional support			
Positive climate	α = .89	α = .91	
Relationships	0.85	0.88	
Positive affect	0.81	0.81	
Positive communication	0.87	0.89	
Respect	0.78	0.83	
Negative climate	α = .72	α = .62	***
Negative affect	0.77	0.70	
Punitive control	0.79	0.67	
Sarcasm/disrespect	0.57	0.55	
Severe negativity	0.41	0.42	
Teacher sensitivity	α = .90	α = .93	
Awareness	0.81	0.89	
Responsiveness	0.89	0.89	
Addresses problems	0.90	0.92	
Student comfort	0.76	0.82	
Regard for student perspectives	α = .82	α = .80	***
Flexability and student focus	0.87	0.72	
Support for autonomy and leadership	0.76	0.62	
Student expression	0.68	0.75	
Restriction of movement	0.62	0.73	
Classroom organization			
Behavior management	α = .87	α = .88	
Clear behavioral expectations	0.70	0.70	
Proactive	0.82	0.80	
Redirection of misbehavior	0.89	0.90	
Student behavior	0.77	0.82	
Productivity	α = .83	α = .88	
Maximizing learning time	0.76	0.80	
Routines	0.82	0.84	
Transitions	0.77	0.80	
Preparation	0.65	0.74	
Instructional learning formats	α = .77	α = .80	***
Effective facilitation	0.77	0.76	
Variety of modalities and materials	0.64	0.64	
Student interest	0.65	0.82	
Clarity of learning objectives	0.68	0.58	
Instructional support			
Concept development	α = .79	α = .79	
Analysis and reasoning	0.73	0.77	
Creating	0.65	0.63	
Integration	0.76	0.71	
Connections to the real world	0.65	0.67	
Quality of feedback	α = .82	α = .80	***
Scaffolding	0.64	0.57	
Feedback loops	0.69	0.58	
Prompting thought processes	0.76	0.73	
Providing information	0.80	0.79	

(continued)

Table 3. (continued)

	Live	Video	*p*
Encouragement and affirmation	0.53	0.63	
Language modeling	α = .78	α = .73	***
Frequent conversation	0.48	0.39	
Open-ended questions	0.76	0.71	
Repetition and extension	0.78	0.71	
Self- and parallel talk	0.56	0.58	
Advanced language	0.72	0.63	

Note. Cronbach's alphas are also presented for each dimension, although those were not a part of the confirmatory factor analyses. Emotional support, classroom organization, and instructional support are the domains. Boldface text indicates dimension. Italicized text indicates indicator.

Is Predictive Validity Similar Across Live and Video Methods?

Finally, to investigate whether there was differential prediction, and therefore differences in the predictive validity of the observation methods, multilevel models were constructed with child academic achievement outcomes. The first step in the analysis was to determine how much variance in each outcome was at the classroom level in the fall and spring. Unconditional models were built that predicted each outcome only accounting for the nesting. Intraclass correlation coefficients (ICCs) quantify the degree of nesting. These analyses (Table 4) indicated that the ICC for TEMA in the fall was .08 and in the spring was .12. For PPVT, the ICC in the fall was .04 and in the spring was .07. For TOPEL, the ICC in the fall was .07 and in the spring was .12.

Conditional models were then constructed that included predictors. These models predicted TEMA, PPVT, and TOPEL outcomes controlling for gender, free and reduced price lunch status, minority status, English language learner status, and baseline score. Individual predictors of interest were added for each model. As CLASS analyses are usually done at the domain level using aggregates (i.e., not factors) and there are moderate correlations between the domains, these analyses included running Emotional Support, Classroom Organization, and Instructional Support as separate predictors, for a total of nine models.

Table 4 shows the results of Emotional Support, Classroom Organization, and Instructional Support separately predicting TEMA, PPVT, and TOPEL. Controlling for gender, free and reduced price lunch status, minority status, English language learner status, and prescore, students, on average, scored a 26.1 on the TEMA, a 45.7 on the PPVT, and a 56.9 on the TOPEL. No clear pattern emerged with either the live or video versions being universally, or even generally, better than the other—each proved to be a better predictor of some outcomes over the other. No domains in either method were predictive of TEMA. For PPVT, domains were predictive for both live and video administrations of the CLASS. However, all three regression coefficients for the video administration were larger, sometimes substantially, than the live versions. However, for the TOPEL, the pattern was somewhat reversed with the live versions of the Emotional Support and Classroom Organization domains being marginally predictive of the outcome, but the video versions not even being marginally predictive. For Instructional Support, the video version predicted the outcome, but the live version was not significantly related at all.

Discussion

The present study focused on comparing and contrasting observations of live and video administrations of the CLASS-PreK (Pianta et al., 2008). Understanding the nature of the differences can help

Table 4. Unstandardized Betas of CLASS Domains Predicting Residualized Change in Child Academic Outcomes.

	TEMA		PPVT		TOPEL	
CLASS domain	Live	Video	Live	Video	Live	Video
Emotional Support	3.66	1.59	3.62*	5.83*	5.19†	3.19
Classroom Organization	1.79	1.56	2.81†	2.98*	4.04†	3.45
Instructional Support	2.17	2.20	3.11*	4.29*	3.10	4.34*

Note. Each analysis above was conducted separately and controls for gender, free and reduced price lunch status, minority status, English language learner status, and baseline score. CLASS = Classroom Assessment Scoring System; TEMA = Test of Early Math Ability; PPVT = Peabody Picture Vocabulary Test; TOPEL = Test of Early Preschool Literacy.
†$p < .10$. *$p < .05$.

inform the users of the CLASS-PreK about the potential benefits of using either live or video method. The present study examined differences between live and video administrations of the CLASS by coding videos that were captured during the live coding. Furthermore, the balancing of raters across the live and video ratings removes the confound that would happen if different pairs of raters were coding different segments. In our examination of the different modalities, several differences were noted in terms of their reliability, means, and ability to predict child academic achievement outcomes.

Reliability Differences

Evident across both modes of coding the CLASS were the sufficiently high alpha coefficients. The one major exception was Negative Climate in the video mode, which was under .70. Given the low frequency of the indicators being observed in this dimension, it suggests that understanding aspects of a classroom's Negative Climate is not best done through video.

Although high alphas are generally seen as a positive aspect of a scale, very high alphas may suggest that some indicators within dimensions are highly redundant. For example, Teacher Sensitivity has alphas above .90 for both live and video codes with only four items. This suggests that the indicators that make up Teacher Sensitivity—Awareness, Responsiveness, Addresses Problems, and Student Comfort—do not add much incremental information over one another. It may also be that the alphas are elevated because of homogeneity in the sample and that other samples may evidence lower alphas. Conversely, scoring at the indicator level for other dimensions, such as Concept Development with alphas of .79, provides some unique information. These alphas suggest that training on Teacher Sensitivity could either be paired down (in the event that the differences are not important to highlight) or increased (to help coders in making distinctions among indicators). Future research could determine whether the unique information is helpful, for example, in predicting child outcomes (cf. Curby & Chavez, 2013).

In the CFAs, similarities and differences were evident in the factor structure by mode of observation. Emotional Support, Classroom Organization, and Instructional Support all showed differences across the live and video factor loadings. At the dimension level, the differences were not due to Positive Climate, Teacher Sensitivity, Behavior Management, Productivity, or Concept Development, which were not shown to be different across modes. However, Negative Climate, Regard for Student Perspectives, Instructional Learning Formats, and Quality of Feedback all have indicators that were structurally different across modes of observation. Thus, certain indicators seem to be more salient in different observation modes.

Mean Differences

There were significant mean differences between CLASS video and live coding scores, such that video codes were generally a little lower than the live codes (fewer observed instances of the targeted behaviors). This finding is consistent with research on the CLASS-Secondary (Pianta et al., 2007), which also generally found lower scores with the video format (Casabianca et al., 2013). These results suggest that live coding allows raters a better vantage point (instances of Negative Climate, Concept Development, Language Modeling), which is reflected as differences in ratings. For example, in the CLASS-PreK, one indicator of Language Modeling is Frequent Conversation (Pianta et al., 2008). It is easily imaginable that live coding allows for better tracking of the teachers' conversation around the room, or even just a better ability to hear the teacher (depending on the technology being used during the video process to capture the audio). Also, our study suggests that Negative Climate may be better observed live than through video (cf. Casabianca et al., 2013), because these scores were higher in the live version, suggesting there were more observable negative behaviors. Future research can determine whether video procedures that provide panoramic views (e.g., v.360 camera) would be helpful in providing scores that are more in line with live codes.

Another possibility for the lower CLASS scores using video was differential rater fatigue. Live coders completed their observations for a classroom in one sitting. Video coders were told to code all the segments for 1 day of a classroom in one sitting (analogous to live coding). However, coders could have completed more than one classroom's videos in 1 day, thus spending more time coding than the live coders. If raters generally provide lower scores as they fatigue (cf. Curby et al., 2011), and they completed more than one classroom in one sitting, then the video scores would be artificially lower. Notably, although the video coding procedure varied slightly from the live coding procedure, it is likely indicative of how video coding may actually take place.

Differences in Predicting Outcomes

Interestingly, trends across the results of the predictive analyses suggest that video and live codes of the CLASS show differential prediction to student academic achievement. CLASS video coding scores seemed to be better at predicting PPVT outcomes with all three domain scores significantly predicting the outcome with larger regression coefficients than live coding. For TOPEL, two of the live coding domains—Emotional Support and Classroom Organization—were marginally predictive, but the video version of Instructional Support was predictive and the live coding version was not. In sum, CLASS live codes significantly predicted two of the nine outcomes and marginally predicted three. CLASS video codes significantly predicted four of the nine outcomes. This does suggest that some findings from previous research linking CLASS scores to outcomes may, in part, be reflecting the coding technique used above and beyond the constructs being measured. It is not known the degree to which differences in coding procedures may account for particular predictive relations using the CLASS in any given study (e.g., Curby et al., 2009; Mashburn et al., 2008), but the present study suggests that there may be differences based on observation modality.

The PPVT is designed as a test of receptive vocabulary whereas the TOPEL is designed to identify students who are at risk of later literacy problems. It could be that the different modalities of coding pick up on differences that are meaningful for these outcomes. For example, a video coder can take breaks or even rewind a video if something was hard to hear or understand. Perhaps the audible vocabulary (heard via video) is more important for the PPVT receptive vocabulary whereas the visible information (seen live) throughout the room is better at predicting TOPEL preliteracy. Finally, it may also be a function of how video cameras were set up. Cameras would often be set up to capture teachers with whole or small groups of children in direct instruction

situations. So, observers would capture those interactions live, but they may not be fully captured on video.

Limitations

The present study provides helpful information about potential differences that exist in live and video coding procedures. However, given its unique set of procedures, several limitations require mentioning. First, the coders used a somewhat unique, but not unheard of, scoring procedure, whereby indicators were scored instead of dimensions. The scores used in this study were used by the schools for assessment of the teachers and classrooms. Thus, this decision is a by-product of the authentic assessment and not a part of the study design. However, it should be noted that aggregated indicator scores were still reliable in the traditional manner.

In addition, teachers may respond to being observed (either video, live, or both) by changing their teaching behavior (Shadish, Cook, & Campbell, 2002) thereby reducing the validity of the ratings. However, because all raters were present with a video camera, the condition was the same for all participants in this study.

Furthermore, it should be noted that there is the potential for carryover effects given that all live observations happened before video observations. Correlations may be inflated by these carryover effects. This concern is mitigated by the fact that there were usually several weeks between when a teacher was observed live and when that coder would have seen videos from that classroom.

Finally, video codes were only coded once. This was intentionally done so that video codes were directly comparable with live codes. However, a great advantage of videos is that they can be scored multiple times. Differences in ratings among observers of the same occasion represent one of the largest sources of variance in CLASS codes (Mashburn, Downer, Rivers, Brackett, & Martinez, 2014), and thus having multiple raters per occasion is likely to yield a more reliable estimate. Future research could compare videos coded multiple times with live codes to see how comparable those results compare to live (and single video) coding.

Conclusion

The present study supports the notion that there are some differences between live and video administrations of the CLASS-PreK. Live codes do offer a couple of advantages in terms of visible (mostly negative) behaviors, presumably because they were happening off screen on the video. Video codes had more significant relations with child outcomes. Given the differences that were present between the two modalities, the present study does suggest that modalities should not be mixed within a study or in assessments of schools. This is particularly true in high-stakes decisions, such as in QRIS evaluations (Tout et al., 2009) when using the CLASS-PreK. The present study suggests that teachers who are observed on video may have artificially lower scores than if those observations were conducted live, meaning that there were fewer observed instances of the targeted behaviors (positive or negative). This is important in light of high profile studies, such as the Measures of Effective Teaching study (Bill and Melinda Gates Foundation, 2012), which examined various classroom observation instruments, all of which were scored based on videos. The present study suggests that results based on video observations may provide somewhat different results than results based on live observations.

Appendix

Results of Chi-Square Change Tests in Multigroup Confirmatory Factor Analyses.

	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*
Emotional support					
Unconstrained	1,770.38	208			
All emotional support indicators constrained	1,948.26	224	177.883	16	<.001
Positive climate constrained	1,779.07	212	8.69	4	*ns*
Negative climate constrained	1,908.23	212	137.85	4	<.001
Teacher sensitivity constrained	1,779.33	212	8.95	4	*ns*
Regard for student perspectives constrained	1,795.78	212	25.40	4	<.001
Classroom organization					
Unconstrained	1,507.75	111			
All classroom organization indicators constrained	1,554.68	123	46.934	12	<.001
Behavior management constrained	1,517.23	115	9.48	4.	*ns*
Productivity constrained	1,512.72	115	4.97	4.	*ns*
Instructional learning formats constrained	1,537.70	115	29.95	4	<.001
Instructional support					
Unconstrained	2,543.77	159			
All instructional support indicators constrained	2,582.30	173	38.532	14	<.001
Concept development constrained	2,546.82	163	3.05	4	*ns*
Quality of feedback constrained	2,561.49	164	17.72	5	<.01
Language modeling constrained	2,562.63	164	18.87	5	<.01

Note. Emotional support, classroom organization, and instructional support are the domains. Boldface text indicates dimension. Italicized text indicates indicator.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was made possible through funding from a U.S. Department of Education Investing in Innovation Grant Award U396C100243 to the AppleTree Institute for Education Innovation.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*, 1034-1037.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Retrieved from http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Arnett, J. (1986). *Caregivers in day care centers: Does training matter?* (Unpublished doctoral dissertation). University of Virginia, Charlottesville.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*, 62-87. doi:10.1080/10627197.2012.715014

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Author.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C., Hamre, B. K. A., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*, 757-783. doi:10.1177/0013164413486987

Curby, T. W., & Chavez, C. (2013). Examining CLASS dimensions as predictors of pre-k children's development of language, literacy, and mathematics. *NHSA Dialog: A Research to Practice Journal*, *16*(2), 1-17.

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R., Howes, C., Burchinal, M., . . . Barbarin, O. (2009). The relations of observed pre-k classroom quality profiles to children's achievement and social competence. *Early Education and Development*, *20*, 346-372. doi:10.1080/10409280802581284

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., . . . Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *The Elementary School Journal*, *112*, 16-37. doi:10.1086/660682

Downer, J. T., Kraft-Sayre, M., & Pianta, R. C. (2009). Ongoing, web-mediated professional development focused on teacher–child interactions: Early childhood educators' usage rates and self-reported satisfaction. *Early Education & Development*, *20*, 321-345.

Dunn, L. M., & Dunn, D. M. (2007). *The Peabody Picture Vocabulary Test, Fourth Edition*. Bloomington, MN: NCS Pearson.

Dunn, L. M., & Dunn, D. M. (2013). *The Peabody Picture Vocabulary Test, Fourth Edition Technical Report*. Retrieved from http://images.pearsonassessments.com/images/assets/ppvt-4/2013-PPVT-Tech-RPT.pdf

Ginsburg, H. P., & Baroody, A. J. (2003). *The test of early mathematics ability* (3rd ed.). Austin, TX: Pro Ed.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *The Early Childhood Environment Rating Scale* (Rev. ed.). New York, NY: Teachers College Press.

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of observation*. Princeton, NJ: Educational Testing Service.

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. (2007). *Test of preschool early literacy*. Austin, TX: Pro-Ed.

Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, *14*, 179-196.

Mashburn, A. J., Downer, J. T., Rivers, S., Brackett, M., & 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*, 146-155. doi:10.1007/s11121-012-0357-3

Mashburn, A. J., Pianta, R., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language and social skills. *Child Development*, *79*, 732-749. doi:10.1111/j.1467-8624.2008.01154.x

Molfese, V. J., Brown, E. T., Adelson, J. L., Beswick, J., Jacobi-Vessels, J., Thomas, L., & Culver, B. (2012). Examining associations between classroom environment and processes and early mathematics performance from pre-kindergarten to kindergarten. *Gifted Children*, *5*(2), Article 2.

National Research Council and National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop* (Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability) (H. Braun, N. Chudowsky, & J. Koenig, Eds.). Washington, DC: The National Academies Press.

Office of Head Start. (2014). *A national overview of grantee CLASS(TM) scores in 2013*. Washington, DC: Office of Head Start, Administration for Children and Families, U.S. Department of Health and Human Services.

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). *Classroom assessment scoring system manual, middle/secondary version*. Charlottesville: University of Virginia.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, NY: Houghton Mifflin.

Teachstone. (2015, October 5). What is the state of CLASS? [Electronic mailing list message]. Retrieved from Marketing Email from interactions@teachstone.com.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of Quality Rating and Improvement Systems* (Issue Brief No. 3). Washington, DC: Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

U.S. Department of Education. (2009). *Race to the top program: Executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Wilson, S. B., & Lonigan, C. J. (2009). Identifying preschool children at risk of later reading difficulties: Evaluation of two emergent literacy screening tools. *Journal of Learning Disabilities*, *43*, 62-76. doi:10.1177/002221940934500