Abstract Title Page

Title: Construct Validity of Classroom Observations: Items, Factors, Raters, and Achievement

Author(s):
Lee Branum-Martin
Coleen D. Carlson
Angie Durand
Christopher Barr

Texas Institute for Measurement, Evaluation, and Statistics (TIMES), University of Houston

Background/context:

Observing classroom literacy instruction is difficult, even under the best conditions. Issues of construct validity for classroom and school quickly become complex if we begin to consider all possible sources of variability which could influence classroom observations, including items, constructs (factors), raters, and time (Raudenbush, 2008; Raudenbush & Sadoff, 2008). However, practical and financial considerations can seriously limit full investigation of all of these facets (Raudenbush, 2008). The current paper represents an attempt to investigate such facets from within a factor analytic approach to construct validity.

Global ratings of instruction and classroom environment represent an important method of measuring educational contexts (Neuman, Koh, & Dwyer, 2008). The Early Language & Literacy Classroom Observation (ELLCO) is an instrument developed to measure instructional behaviors and environmental conditions conducive to early literacy (Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002). The ELLCO has been found to be sensitive to changing teacher behaviors as a result of professional development (Grace, et al., 2008; Gray, 2007), but not always (Ball & Gettinger, 2009). The ELLCO has also been found to be related to student reading fluency (Ball & Gettinger, 2009) and possibly indicative of the mediating influence of classroom environment for literacy behaviors (Wayne, DiCarlo, Burts, & Benedict, 2007).

The ELLCO was designed to measure several facets of classrooms and instruction, including organization, management, climate, and instruction (see Table 3 for facets used in this study). Observers make a global rating of the quality of each facet, based on descriptive behavioral anchors for the quality of support to language and literacy. Aside from interrater reliability and internal consistency, no empirical models of the construct validity of the ELLCO have been found. Therefore, the current study split these global judgments into items specific to particular behaviors and classroom characteristics so that the validity of measurement could be investigated via a confirmatory factor analysis model (CFA) of items to intended constructs (facets of the classroom).

Purpose / objective / research question / focus of study:

The current research sought to examine the construct validity of an observational instrument with a confirmatory latent variable model, to examine the reliability and stability of derived factor scores, and to examine their relation to school-level academic performance. In this way, we sought to examine the validity of an empirically based a priori model of the effects of the classroom environment.

Setting:

In each year, the evaluation of Texas Reading First selected schools for observations in order to measure classroom instruction. In the last 3 years, observations have been conducted at 158 schools across Texas.

Population / Participants / Subjects:

The 2,172 teachers who were observed over the three years were 93% female and 45% Hispanic, 41% White, and 11% African American. In terms of education, 82% held a Bachelor's

degree and 16% had a Master's or doctorate. 25% of the teachers held a certification in Early Childhood Education and 1% were credentialed Master Reading Teachers. They had been teaching for an average of 10 years (SD = 9.5), but half had taught for 7 or fewer years and the modal experience was 2 years. Table 1 shows the number of observations made per year, semester, and grade level.

Intervention / Program / Practice:

The evaluation was designed to measure the impact of funding at the campus level, so a sampling approach was designed to characterize instruction for each grade level at the campus: at least 2 teachers per grade level per campus. Observers were former classroom teachers who were trained to use the observation instrument. Observers were trained for 16 hours each semester and practiced on the instrument until they exceeded 90% agreement on videotaped lessons.

Research Design:

In school years 2006-2007 and 2007-2008, a random subsample of Reading First eligible campuses was selected for observations because funding was not available to observe all 800+ campuses. In school year 2008-2009, funding was only made available to observe 100 campuses. For the purposes of investigating the construct validity of the observation system, the sample is a quasi-experimental group of low-performing Title I schools.

Data Collection and Analysis:

At each selected campus, two teachers were randomly sampled to be observed for 60 min of their required 90 min reading and language arts instruction block. The instrument consisted of 21 items designed to measure 8 aspects of instruction and the classroom environment (Table 3). Each item was rated on a 3-point scale, and each response point had a specific description for the quality of support to student language and literacy. For the current purpose, we will refer to these response points as "high," "medium," and "low" rated quality. Examples of the Management and Climate items are shown in Table 7.

Using a restricted factor model developed in school year 06-07, we fit this model in a confirmatory manner to all years and semesters of data. The model is a CFA for ordinal items which has a simple structure for the factors, each measured by 3-category items. The structure of items to factors is shown in Table 3. All ordinal CFA models were fit with robust weighted least squares estimation in Mplus 5.2 (WLSMV; Muthén & Muthén, 2007). The analysis followed 7 steps:

1. Fit CFA model separately to each year, grade, and semester.
2. Fit CFA model across grade, within year and semester.
3. Fit CFA across grade and year, within semester (one model each for fall and spring).
4. Fit one CFA across grade, year, and semester. Apply factor loadings from this model to score all observations. In this way, each factor is scored on a metric which is consistent across grades, years, and semesters.
5. Examine reliability of observers on factor scores.
6. Examine stability of scores within teacher and school.
7. Examine relation of school mean factor scores with school mean achievement.

Findings / Results:

The results from Step 1 indicated that the CFA for each year, grade, and semester was quite reasonable (CFI/TLI near .95, RMSEA < .08). The only exception to the model was that there was not sufficient writing instruction observed in kindergarten, so the writing items were excluded from kindergarten-only models.

Because the models from Step 1 were reasonable, across-grade models were fit for each year and semester. The fit statistics from these models are presented in Table 2. In all, the models fit reasonably. The models in later years (07-08 and 08-09) had some degree of mis-fit (CFI < .95; RMSEA > .08), but were not unreasonable. Taken together, these models suggested that a joint model for each semester might be tenable.

Step 3 tested these semester models across years and grades. These models fit reasonably (Table 2) and showed that despite minor shifts in a few items, the overall measurement relations still conformed to the a priori model.

In step 4, the results from the spring model in Step 3 were used to fit a single model across all years, grades, and semesters. This scoring model had all factor loadings thresholds fixed to the values estimated in the step 3 spring model. In this way, the factor means and variances were identified on an across-grade, across-year metric, allowing us to make inferences about changes in the latent constructs. Despite the strong restrictions, this confirmatory model had excellent fit (Table 2).

Table 3 shows the standardized factor loadings, loading SE, and thresholds for each item by factor. All loadings were statistically significant ($p < .001$). Thresholds represent boundaries between the 3 categories of quality for each item. The first threshold (t1) between "low" and "medium" quality for most items fell 1 SD or more below the mean on their respective factors, indicating that classrooms were rarely rated as being poor. The threshold between "medium" and "high" for most items fell near the mean (zero) of their respective factors. Figure 1 shows the item thresholds on their factor-specific z-scales.

Table 4 shows the correlations among estimated factor scores. The correlations were positive, but not homogeneous ($r = .27$ to $.79$).

In step 5, the factor scores were then examined for reliability in two ways: interobserver agreement and intraclass correlation. First, a random subset of 68 observations were made simultaneously by two observers on actual lessons being conducted in the schools. In no case was there a significant mean difference, and all correlations on the factor scores were .70 or above, except for Organization (interrater $r = .49$).

Second, because observations were nested within observer, variance components were estimated in a multilevel model of 4,020 observations cross-classified by 2,173 teachers and 84 observers across the three years of the study (fit in SAS PROC MIXED; Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Results indicated that 13-24% of the variability was across teachers and 19-25% of the variability was due to the observer. Additional three-level models including schools generally yielded less than 5% variability across campuses (not reported here).

In step 6, we examined the stability of factor scores over time within teacher and school. In regressing each of the factor scores on time resulted in no model with an R-squared value greater than .007, indicating no linear relation over time for teachers. Visual inspection of school average factor scores showed randomly crossing lines, indicating essentially no temporal stability in school instructional quality.

In step 7, scatterplots were examined between campus mean factor scores and campus mean reading achievement. Most of them were reasonably linear, but there was substantial fluctuation in the relation across years and semesters. That is, the relation between instruction and achievement at the campus level was usually, but not always, positive. Table 6 was constructed to give an overall impression of the degree of linear association between campus instruction and achievement for 3 of the 7 factor scores.

Conclusions:

The application of this observation instrument, like many used for instruction, involves categorical items nested within factors, factors nested within observation times, observations within teachers, teachers within (and crossed with) observers, and teachers within schools. Any one of these sources of variability can jeopardize the inferences we wish to make about the quality of actual instruction delivered (Raudenbush, 2008).

The current results highlight a number of important points. Ordinal SEM models can be applied as CFA for a priori instrument validation. SEM provides a framework for investigating validity in a number of ways, from configural invariance as tested here, to full investigation of item parameters at each time point (not estimable here). The results represent how an item response approach can yield validity information about observations of instructional behavior. The factor model from the first year of implementation was replicated across 2 semesters each in 2 subsequent years, for a total of 6 waves of measurement. This replication provided several opportunities for falsification of the intended model.

It is possible that alternative models could also be tried, but the purpose of the current investigation was an a priori replication of the model intended by developing the observation instrument. Visual inspection of Table 4 suggests that the instructional factors of Reading and Writing may be different from the other factors which appear more homogeneous. While the current results include 8 factors, their interpretation is relatively straightforward, with no cross-loadings or conceptually mixed items.

The item loadings indicated strong measurement of their intended factors. The item thresholds indicated very few responses in the low category, and up to half or more in the "high" category (i.e., thresholds greater than zero). This may indicate that the instrument used is more sensitive in distinguishing low performing teachers than high performing teachers. While it is possible that social desirability prevented observers from giving many ratings below "high," the high responses may suggest a need to revise the behavioral anchors in the ratings to create a fourth category, perhaps for "excellent." The current results improve upon simple item descriptive statistics since the factor to item correlations support construct validity, but the thresholds suggest room for making the items more sensitive to distinctions among "high" performing teachers. Figure 1 is diagnostic in this regard for the location of item categories on the underlying classroom factors.

The reliability investigation revealed that while factor level correlations were high and no mean differences were found between raters, approximately 20% of the variability in factor scores was due to raters. Few reliability observations were possible in the current project, but more attention should be given to this area. Assignment of raters to teachers was made on practical considerations regarding travel across the state, so while three-level models including school did not change the amount of variance attributable to raters, systematic differences due to regions or locations may require further investigation.

While the conventional correlations and mean differences (on estimated factors) revealed no serious concerns, current results suggest that interrater variability may be a large and expensive concern to investigate.

With regard to stability over time, factor scores exhibited little stability within teacher, perhaps reflecting the campus-level sampling scheme of twice per year. Campus level stability was not very high, either, which may reflect high teacher turnover, but is also a question for a more detailed longitudinal investigation at the teacher level.

External validity was investigated via correlations with school-level achievement. At the campus level, factor scores had low to moderate correlations with campus reading achievement. It is important to note that aside from the reading instruction factor, relations to school level performance are not a strict indicator of criterion validity. Such relations might not be expected to be strong, given the aggregation across teachers to the campus level, as well as that the observations targeted general classroom quality (albeit in specific facets of classroom activity). Moreover, caution must be exercised, since schools vary widely in size and grade levels served, so the likelihood of influential or outlying values is high. These school level results could be made stronger with direct linkage of student identifiers with teacher identifiers provided by the test publishers, allowing investigation at the classroom level (however, the provided data did not always have the proper classroom teacher identified). These relations suggest that campuses with better classroom instruction and environments have better reading achievement, but there is a considerable amount of fluctuation between semesters and across years. As given, the current results provide a rigorous basis for examining campus-level instructional quality across grades and across time.

Despite the daunting task of multifaceted measurement under practical and financial constraints, reasonably rigorous measurement is possible. The current results suggest promise for further examination for their instructional implications. Item response models, in the form of ordinal SEM, provide a method of investigating the construct validity of observational data. In the current study, items and factors were found to conform to expected relations. However, raters and schools over time were found to have surprising amounts of variability. Variability across raters and time deserves closer examination in the same way that the nesting of students within schools has received (Hedges & Hedberg, 2007) in order to properly delimit the generalizability of classroom observation systems.

.

Appendices

Appendix A. References

Ball, C., & Gettinger, M. (2009). Monitoring children's growth in early literacy skills: Effects of feedback on performance and classroom environments. *Education & Treatment of Children, 32*(2), 189-212.

Grace, C., Bordelon, D., Cooper, P., Kazelskis, R., Reeves, C., & Thames, D. G. (2008). Impact of professional development on the literacy environments of preschool classrooms. *Journal of Research in Childhood Education, 23*(1), 52.

Gray, S. (2007). Evaluation of a program to promote early literacy skills in preschool children. *Early Childhood Services: An Interdisciplinary Journal of Effectiveness, 1*(1), 17-31.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Littell, R. D., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide. Fifth edition.* Los Angeles, CA: Muthén & Muthén.

Neuman, S. B., Koh, S., & Dwyer, J. (2008). CHELLO: The Child/Home Environmental Language and Literacy Observation. *Early Childhood Research Quarterly, 23*(2), 159-172.

Raudenbush, S. W. (2008). *Statistical inference when classroom causality is measured with error*. Paper presented at the Society for Research on Educational Effectiveness.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*(2), 138-154.

Smith, M. W., Dickinson, D. K., Sangeorge, A., & Anastasopoulos, L. (2002). *Early Language & Literacy Classroom Observation (ELLCO) Toolkit, Research Edition*. Baltimore, MD: Paul H. Brookes.

Wayne, A., DiCarlo, C. F., Burts, D. C., & Benedict, J. (2007). Increasing the literacy behaviors of preschool children through environmental modification and teacher mediation. *Journal of Research in Childhood Education, 22*(1), 5.

Appendix B. Tables and Figures

Table 1: Number of observations made per year and semester by grade level

Year	Semester	Grade K	Grade 1	Grade 2	Grade 3	Total	Unique Teachers	Unique Schools
06-07	Fall	219	216	210	205	850	850	117
	Spring	197	207	199	203	806	800	106
07-08	Fall	198	205	196	185	784	776	102
	Spring	193	214	202	191	800	767	102
08-09	Fall	102	101	102	95	400	378	50
	Spring	97	100	94	84	375	375	50
	Total	1,006	1,043	1,003	963	4,015	2,172	158

Table 2: Fit statistics from the models in analysis Steps 2-4.

Step	Model	χ^2	*df*	CFI	TLI	RMSEA	WRMR
2	06-07 Fall	317.8	89	.958	.980	.055	1.126
	06-07 Spring	290.0	89	.959	.980	.053	1.090
	07-08 Fall	434.2	84	.925	.955	.073	1.418
	07-08 Spring	363.5	84	.937	.960	.066	1.310
	08-09 Fall[a]	263.1	64	.926	.957	.091	1.212
	08-09 Spring	202.2	59	.952	.968	.080	1.114
3	Fall across years, grades	768.8	94	.943	.971	.060	1.714
	Spring across years, grades	661.8	93	.951	.973	.056	1.622
4	Scoring all waves[b]	849.4	77	.968	.979	.050	2.315

Notes: All Chi-square statistics were significant ($p < .001$). [a] For the Fall 08-09 model, the output contained warnings regarding two observed correlations near unity and a non positive definite latent variable matrix, involving the writing factor (but no inadmissible values were found). [b] All loadings and thresholds were fixed to the values estimated in the step 3 spring model. CFI = Comparative Fit index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Squared Error of Approximation. WRMR = Weighted Root Mean Square Residual.

Table 3: A priori factors, items, and their model estimates

Classroom Factor	Item	Loading	SE	t1	t2
Assessment	Range of techniques	.85	.01	-1.11	0.42
	Interactions	.90	.01	-1.74	0.16
Climate	Tone of conversations	.94	.02	-1.91	-0.47
	Listening	.94	.02	-2.43	-0.44
Curriculum Integration	Goals	—	—	-1.86	-0.59
Management	Rules	.77	.02	-1.72	-0.23
	Expectations	.92	.04	-1.96	-1.05
	Conflict intervention*	—	—	—	—
Oral Language	Aware of development	.48	.03	-1.89	-0.29
	Conversation	.84	.01	-1.18	0.16
	Uses	.88	.01	-1.07	0.12
	Vocabulary	.77	.02	-1.51	-0.27
Organization	Size of furnishings	.62	.02	-2.74	-1.27
	Arrangement	.84	.02	-1.85	-0.81
	Engagement	.90	.02	-2.15	-0.92
Reading Instruction	Experiences	.70	.02	-1.17	-0.55
	Comprehension	.85	.02	-0.92	-0.03
	Instruction	.70	.02	-1.13	0.11
Writing Instruction	Opportunities	.77	.04	-0.87	0.14
	Materials	.77	.03	-0.91	-0.09
	Variety	.86	.03	-0.75	-0.15

Notes: t1 = threshold between low-medium on factor-specific z-scales. T2 = threshold between medium-high on factor-specific z-scales. The "Goals" item was a single indicator of the curriculum integration factor, and therefore is taken as the only representation of the factor (no loading). The conflict intervention item was dropped because covariance could not be estimated with listening and expectations items.

Table 4: Factor correlation matrix

Factor	Assess	Climate	Curric.	Man.	Oral	Org.	Read.	Writ.
Assessment	1.00							
Climate	.53	1.00						
Curriculum	.58	.55	1.00					
Management	.64	.56	.62	1.00				
Oral Language	.61	.77	.63	.63	1.00			
Organization	.63	.56	.59	.59	.79	1.00		
Reading	.45	.48	.39	.53	.49	.52	1.00	
Writing	.37	.49	.55	.57	.27	.37	.40	1.00

Notes: All correlations were statistically significant ($p < .001$).

Table 5: Interobserver agreement on estimated factor scores

		Observer		Comparison			Teacher	Observer
Factor	Correlation	mean	SD	mean	SD	*p*	ICC	ICC
Assessment	.76	.02	.65	-.06	.76	.43	.13	.23
Climate	.81	-.09	.75	-.15	.78	.33	.21	.19
Management	.73	-.08	.56	-.09	.57	.99	.24	.19
Oral Language	.84	.01	.42	-.05	.47	.11	.19	.20
Organization	.49	-.08	.38	-.02	.34	.19	.14	.25
Reading	.86	-.01	.59	-.08	.67	.30	.17	.18
Writing	.70	-.09	.51	-.17	.54	.21	.14	.20

Note: n = 68 for the interobserver measures. "Observer" refers to the factor score for the observer. "Comparison" refers to the score derived for the trainer who was sent as the comparison observer. The *p*-values are for paired samples *t*-tests between the observer and comparison. ICC = intraclass correlation for a model of 4,015 observations cross-classified by 2,172 teachers and 84 observers across the three years of the study.

Table 6: Correlations between fall and spring campus level mean factor scores and campus level reading achievement

				Reading Instruction		Oral Language Instruction		Classroom Management	
Test	Grade	Year	Schools	Fall	Spring	Fall	Spring	Fall	Spring
ITBS	1	06-07	78	.21	.38*	.20	.39*	.24*	.38*
		07-08	68	.12	-.05	.17	-.05	.23	.03
		08-09	33	.01	.14	-.06	.10	-.16	.09
	2	06-07	78	.06	.37*	.13	.36*	.26*	.31*
		07-08	69	.12	-.01	.16	.06	.33*	.11
		08-09	33	.39*	.57*	.39*	.56*	.22	.54*
SAT-10	1	06-07	30	.24	.16	.18	.32	.10	.39*
		07-08	29	.34	.06	.34	.03	.24	-.12
		08-09	15	-.01	.07	.12	.07	.17	.30
	2	06-07	30	.09	.36*	.14	.32	.13	.31
		07-08	29	-.03	.03	.02	-.00	.02	-.23
		08-09	15	.21	.30	.23	.17	.37	-.06
TAKS	3	06-07	105	.22*	.21*	.26*	.18	.31*	.31*
		07-08	96	.13	.19	.08	.16	.15	.22*
		08-09	46	.01	.20	.03	.11	.14	.33*

Notes: ITBS = Iowa Test of Basic Skills. SAT-10 = Stanford Achievement Test-10. TAKS = Texas Assessment of Knowledge and Skills.

Table 7: Examples of Classroom Management and Climate Items

Classroom Management

Understanding of rules

1) Children appear to have limited understanding of rules and routines. This is evident in the classroom as children engage frequently in conflicts and rarely in purposeful activity.
2) Children appear to understand regular rules and routines, but there is occasionally a need for teacher reminders or reinforcement about some rules and routines.
3) Children appear to have internalized regular rules and routines. This is evident as children move through the classroom period smoothly, with few conflicts, and are most often seen engaged in purposeful activity.

Communication of expectations

1) Expectations for children's behavior may be confusing or inconsistent, conflicts may be inconsistently resolved.
2) Expectations for children's behavior are communicated from teacher to children.
3) Clear expectations for children's behavior are consistently communicated in multiple ways.

Conflict management

1) The teacher may fail to identify conflicts or may resolve them in an arbitrary or harsh manner.
2) The teacher consistently resolves conflicts for children. For example, teachers may separate certain children or may provide alternative materials without encouraging children to share or take turns.
3) Teacher intervention in conflicts is calm, non-threatening, and leads children toward peaceful, independent (i.e., alone or with peers) resolutions.

Classroom Climate

Tone

1) The tone of classroom conversations may be negative, or the teacher's manner may be harsh or punitive. Alternatively, the teacher may appear "distant" or "tuned out" and unavailable to children.
2) The tone of teacher-child conversations in generally positive. Teachers engage in conversations with children but do not typically encourage voicing of multiple and diverse perspectives.
3) The tone of classroom conversations is positive and shows respect for children's contributions, encouraging children to speak from their different perspectives and experiences.

Listening

1) Children are expected to listen to the teacher and there are few opportunities for conversation.
2) Teachers listen to children but do not intentionally encourage children's conversations with each other. Similarity and convergence of opinions are valued.
3) Teachers listen attentively to children, encourage children to listen to each other, and deliberately foster a climate in which differing opinions and ideas are valued.

Figure 1: Item thresholds between rating categories on their factor-specific z-scales (diamond = low to medium; square = medium to high)