A Study of the Predictive Validity of the LSC Classroom Observation Protocol

Eric R. Banilower

April 2005

Submitted to: The National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230

Submitted by: Horizon Research, Inc.
326 Cloister Court
Chapel Hill, NC 27514

Introduction

One of the key components of the core evaluation of the Local Systemic Change through Teacher Enhancement Initiative (LSC) is the observation of a random sample of lessons. Data from these classroom observations, along with other core evaluation data, are used to evaluate the impact of the LSC on the nature and quality of classroom instruction. In addition, the Classroom Observation Protocol (COP), developed for the LSC core evaluation, is widely used for non-LSC evaluations and as a professional development tool. Although review by experts in mathematics/science education have provided a degree of face validity to the COP, no formal study has been conducted to determine whether ratings of lesson quality using the COP predict student learning gains. Thus, a research study was designed to examine whether this link exists.

Study Design

The purpose of the study is to provide evidence of the validity of the LSC COP, by correlating observers' ratings of lessons to changes in student achievement. The study design called for observing 40 mathematics classes three times each over the course of a unit. In order to control for as many variables as possible, the study was to be limited to a single grade level and curriculum unit.

HRI worked with the district mathematics leadership team to select the targeted grade level and curriculum unit. Fourth grade and the *Landmarks in the Thousands* unit from the *Investigations* curriculum were selected for a number of reasons. First, the content of the unit was well aligned with the district mid-year assessment used at that grade level. Second, the district pacing guide called for the teaching of this unit just before the mid-year assessment. Third, the district mathematics leadership team thought that more fourth grade teachers were implementing the *Investigations* curriculum than teachers in other grade levels.

HRI planned to select 20 matched pairs of teachers for the study. Each pair would be comprised of a strong implementer and an emerging implementer of the *Investigations* curriculum (as judged by the district mathematics leadership team, with neither the observers nor the investigators being aware of a teacher's designation). Teachers would be matched on factors such as teaching experience and school demographics, thus controlling for other factors known to affect student achievement. HRI planned to offer a $150 honorarium to each teacher participating in the study.

The study design called for observing three lessons within the instructional unit observed by a trained researcher. In addition, teachers would participate in a 30-minute interview following each lesson, following which the researcher would complete a COP and assign a quality rating to the lesson.

Student scores on the district mid-year mathematics assessment would be analyzed in respect to the observational ratings using multilevel regression. Prior achievement and various student demographic characteristics—gender, race/ethnicity, free/reduced-price lunch status, English

proficiency status, and whether the student has an individualized education plan would be included in the model as covariates.

However, even with the $150 honorarium for participating in the study, HRI encountered great difficulty in recruiting teachers for this study. In addition, during the fall semester, the district decided to do away with its mid-year mathematics assessment. In consultation with the district mathematics leadership team, HRI decided to broaden the study to include 3rd and 5th grade teachers, in addition to 4th grade teachers, with the sole requirement being that they were using the *Investigations* curriculum. The outcome measure to be used was switched to the state end-of-year assessment.

Still, despite repeated recruitment attempts (both by fliers and in-person by members of the district mathematics leadership team), only 18 teachers participated in the study. The school district was able to provide student data for 16 of the 18 classes, further reducing the sample size of the study.

The reduced sample size had a dramatic effect on the study's statistical power. Rather than the 80 percent chance of detecting an effect of 0.35 standard deviations as originally planned, a post-hoc power analysis indicated that the actual sample provided only a 40 percent probability of detecting an effect of this size and a 66 percent probability of detecting an effect of 0.5 standard deviations.[1] Thus, unless the relationship between ratings on the COP and student achievement scores is very strong, it is unlikely that this study would detect that relationship.

The Sample

Twelve of the 16 teachers in the final sample taught 4th grade (see Table 1), which is most likely a reflection of the initial targeting of that grade level. Three teachers taught 5th grade and one taught 3rd grade.

Table 1
Grade Level Taught

	Number of Teachers (N = 16)
Grade 3	1
Grade 4	12
Grade 5	3

Of the 352 students enrolled in these teachers' classes at the end of the year, HRI received complete data for 309 students. Table 2 provides demographic data for these students. Roughly half of the students were female; nearly two-thirds were a member of a non-Asian minority group. Forty-five percent of the students were eligible for free/reduced-price lunch (FRL).

[1] Power calculations made using Optimal Design software by Raudenbush, S.W., Spybrook, J., Liu, X., & Congdon, R. (2004)

Thirty-one percent were classified as academically or intellectually gifted (AIG); 14 percent had an individualized education plan (IEP) for a physical, mental, or emotional disability. Six percent of the students were classified as limited English proficient (LEP).

Table 2
Student Demographics

	Percent of Students (N = 309)
Gender	
Female	48
Male	52
Race/Ethnicity	
Non-Asian Minority	63
White/Asian	37
Grade Level	
Grade 3	7
Grade 4	72
Grade 5	21
Eligible for Free/Reduced-Price Lunch (FRL)	45
Classified as Limited English Proficient (LEP)	6
Classified as Physically, Mentally, or Emotionally Disabled (IEP)	14
Classified as Academically or Intellectually Gifted (AIG)	31

Analysis and Results

The data used in this study have a nested structure, with students nested within teachers' classrooms. Statistical techniques that do not account for potential shared variance within groups in nested data structures can lead to incorrect estimates of the relationship between independent factors and the outcome. Hierarchical modeling is an appropriate technique for apportioning and predicting variance within and across groups in a nested data structure (Bryk & Raudenbush, 1992).

A two-level hierarchical linear model (students nested within teachers' classrooms) was used to investigate the relationship between ratings on the COP and student achievement. In addition, a number of student demographic factors were controlled for in the model, for example, gender, FRL status.

The independent variables included at the student level were:

- Prior achievement (measured by the previous year's state test score);
- Gender (female vs. male);
- Race/ethnicity (non-Asian minority vs. Caucasian/Asian);
- Eligible for free/reduced-price lunch (FRL);
- Classified as limited-English proficient (LEP);
- Classified as physically, mentally, or emotionally disabled (IEP); and
- Classified as academically or intellectually gifted.

At the teacher level, the independent variables were:

- Grade level (dummy coded); and
- Average observation rating.

Ratings of the teachers' lessons were measured on the seven-point COP scale. As can be seen in Table 3, the majority of lessons were rated on the lower end of the COP scale; fewer than 20 percent of the lessons would be considered high quality (a rating of Level 3 high, 4, or 5).

Table 3
Distribution of Lesson Ratings

	Percent of Lessons (N = 48)
1: Level 1	6
2: Level 2	40
3: Level 3 low	19
4: Level 3 solid	17
5: Level 3 high	4
6: Level 4	15
7: Level 5	0

The three observation ratings for each teacher were averaged together to create a single quality rating. As can be seen in Table 4, the lowest rating for a teacher was under 2, the highest was just over 5 (a "3 high"). The mean observation rating across teachers was 3.17, just over a rating of "3 low."

Table 4
Mean Observations Ratings for Teachers

	N	Minimum	Maximum	Mean	Standard Deviation
Mean Observation Rating	16	1.67	5.33	3.17	1.33

Student assessment scores were standardized within grade level using state-wide means and standard deviations. Table 5 provides descriptive statistics for both the standardized outcome and prior achievement scores. For the entire sample, the mean scores at both time points are essentially the same as the state-wide mean. The intra-class correlation coefficient for the outcome is 0.31; in other words, 31 percent of the variance in the outcome is between teachers.

Table 5
Standardized Student Achievement Scores

	N	Minimum	Maximum	Mean	Standard Deviation
Prior achievement	309	-3.35	2.81	0.02	1.09
Outcome	309	-2.92	2.48	0.05	1.09

HLM 6.00[2] was used for the analysis, with all independent variables entered using grand-mean centering. In addition, the Level 2 random effects were tested to determine if the relationship between the student level (Level 1) predictor variables and the outcome variable varied across teachers. In cases where these effects did vary, the teacher level variables were used to predict this variation. The final estimates of the regression coefficients and their standard errors are shown in Table 6.

Table 6
Regression Coefficients and Standard Errors

	Student Achievement
Intercept	0.07 (0.05)
Average observation rating	-0.05 (0.05)
Grade Level (4th Grade omitted)	
3rd Grade Class	0.22 (0.23)
5th Grade Class	0.11 (0.14)
Student prior achievement	0.69*** (0.04)
Non-Asian minority student	-0.20* (0.08)
Female student	0.04 (0.06)
AIG student	0.20* (0.09)
FRL student	-0.11 (0.08)
LEP student	-0.05 (0.13)
IEP student	-0.24* (0.11)
Average observation rating	-0.03 (0.10)
Grade Level (4th Grade omitted)	
3rd Grade Class	-0.88** (0.28)
5th Grade Class	0.47~ (0.25)

~ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

[2] Raudenbush, S., Bryk, A., & Congdon, R. Scientific Software International, 2004.

The regression analysis does not provide evidence of a relationship between student achievement scores and the average COP rating, controlling for prior achievement and demographics. This finding is not surprising given the low statistical power of the study. As noted earlier, the final sample size provided only a 40 percent chance of detecting an effect of 0.35 standard deviations.

As would be expected, prior student achievement was a significant predictor of current achievement. Additionally, non-Asian minority students scored, on average, 0.20 standard deviations lower than Caucasian/Asian students. Students classified as AIG typically scored 0.20 standard deviations higher than non-AIG students. None of these relationships varied significantly across teachers.

Students with an IEP tended to score lower than students without an IEP (0.24 standard deviations), though this effect varied among teachers. IEP students in 3rd grade tended to do worse and IEP students in 5th grade tended to do better than IEP students in 4th grade. In other words, the difference in scores between IEP and non-IEP students tended to narrow in the higher grades. The average observation rating from the COP was not related to IEP student performance, though the small number of teachers in the sample makes finding such a relationship unlikely. Including grade level and average observation rating explained all of the across-teacher variation in the IEP-achievement score relationship.

Conclusions

The purpose of this study was to provide additional data for evaluating the validity of the LSC Classroom Observation Protocol (COP). Although face validity of the COP has been established through expert review, no other studies of the COP's validity have been undertaken. This study looked at the predictive validity of the COP by examining the extent to which observer ratings of the quality of classroom instruction using the COP was correlated with student learning gains.

This study examined the teaching and learning of elementary grades mathematics in a single school district. Unfortunately, the number of teachers participating in the study was much lower than planned, resulting in a study with limited statistical power (i.e., a low probability of detecting an association should it truly exist). Perhaps due to the low statistical power, the results of this study do not provide evidence of the predictive validity of the COP. Controlling for prior achievement and student demographics, observer ratings of the quality of instruction using the COP were not significantly related to student assessment scores.

It is important to note that although this study does not provide evidence of the predictive validity of the COP, it also does not provide evidence of the lack of predictive validity of the COP. The small number of teachers participating in the study severely limits the ability to draw any conclusions from these data. In addition, the COP was intended to be used to evaluate science and mathematics teaching in grades K–12. A comprehensive study, or set of studies, of the predictive validity of the COP would examine teaching and learning in a number of locations across both subject areas and the entire grade span, an undertaking that was not possible at this time.