Supplementary Material for Manuscript

Effect of Observation Mode on Measures of Secondary Mathematics Teaching

Jodi M. Casabianca

Carnegie Mellon University & RAND Corporation

Daniel F. McCaffrey

Educational Testing Service

Drew H. Gitomer

Rutgers, The State University of New Jersey

Courtney A. Bell

Educational Testing Service

Bridget K. Hamre

University of Virginia

Robert C. Pianta

University of Virginia

Table of Contents for Supplementary Material

Data Collection Process

In the following we describe the data collection process for the *Towards an Understanding of Classroom Context* (TUCC) study that explores best practices for implementing the Classroom Assessment Scoring – Secondary (CLASS-S) instrument.

Raters and Training

In the summer of 2009, we recruited seven school researchers (SR), from 86 applicants, to conduct observations in the schools, collect video, and score the videos. Five SRs completed training and demonstrated the level of professionalism required of the position. All five SRs had been secondary public school teachers for at least two years. Four SRs had experience in the state where the study took place and two had previously worked in the school district. Four of the SRs had experience teaching secondary mathematics and the fifth SR taught English/Language Arts. The group was made up of one white and four African-American women.

The school researchers underwent six days of training in preparation for their work in the study. This training included learning how to use the CLASS-S observation protocol as well as how to conduct all study procedures. The SR's CLASS-S training was two full days long, and was conducted by the developers of the CLASS-S instrument. After the training days, SRs studied the CLASS-S materials and then completed a reliability test. The test consisted of five 15 minutes videos in which the SRs were required to accurately assign scores such that their scores agreed with 80% of the ratings assigned by the CLASS-S master raters. Further, if the SRs made errors they had to be distributed across the various domains of the protocol. This ensured that SRs understood the three domains of CLASS-S with similar levels of accuracy.

The additional four days of training taught the SRs how to handle all other study procedures including: recruitment activities, video camera and computer use, study-specific

software packages, standards for professional and ethical research conduct, and all administrative study procedures.

Assignment of Raters

To assign SRs to the observation of classes, we created a rater assignment table that assigned raters to the four observation lessons for each classroom. The rows corresponded to groups of classrooms and the columns to the observation lessons. Each row contained four cells with one or two rater identifiers of the rater or raters who would observe the classroom. We made the actual assignments by randomly assigning a classroom to each row and an SR to each rater identifier.

The rows of the assignment table were blocked into days so that all rater identifiers had an assignment on every day. Within a row, three lessons were assigned one rater and the other lesson received two raters for three quarters of the rows (75% of the classrooms) and one rater for one quarter of the rows (25% of the classrooms). We made the assignments so that each lesson received two raters for 25 percent of rows. Hence, the first lesson received two raters for 25 percent of the assignment rows, the second lesson received two raters for 25 percent of the assignment rows, and so on. Overall, 20 percent of cells in the table received two raters, so that 20 percent of lessons would be double scored. For rows without a doubled scored lesson, we assigned three raters, two to observe one lesson each and a third to observe two lessons. For rows with a double-scored lesson, three raters were assigned to observe one lesson each and a fourth rater was assigned to observe two lessons. Table 1 shows the assignment of raters using a scenario with 10 classrooms. The table includes assignments for the six SRs participating in the study when we created the original assignments. According to the table, the first lesson observed for the group of classrooms assigned to row one will be observed by the SR randomly assigned

Table 1

Example Rater Assignment Table for Live Observations.

	Observation Period			
Classroom	1	2	3	4
1	B	CA	D	B
2	CA	D	E	C
3	D	E	FA	D
4	E	F	B	EA
5	F	B	C	F
6	AB	D	E	A
7	C	EB	F	C
8	D	F	A	DB
9	E	A	CB	E
10	F	C	D	F

Note. The table demonstrates the assignments of raters to classrooms. The labels A – F designate the raters in the project. Eight of the 40 lessons (20%) are double scored. The percentages of lessons receiving two raters is not 25% in this example rater assignment table since we only present two days in the schedule of observations.

to be rater B. The lesson from the second observation lesson for this group of classrooms will be observed by the SRs assigned to be raters A and C. The lesson for the third lesson will be observed the SR assigned to rater D and the lesson for the fourth lesson will be observed by the SR assigned to rater B. The first five rows correspond to the first day grouping and for each lesson all the SRs have an assignment. For example on the first lesson, rater B will observe classroom group 1, raters C and A will observe classroom group 2, rater D will observe classroom group 3, rater E will observe classroom group 4, and rater F will observe classroom group 5. On the second, third, and fourth lessons every rater is again assigned a group of classrooms. The plan was that on the first day of observations the SRs would each conduct the

observations according to the first day schedule so that all the SRs were busy every day and the required double scores could occur without conflicts. Similarly on the second day of the first round of observations the SR would follow the day 2 plan from column one and so on until all the first round observations occurred. On the first day of the second round of observations, the SRs would follow the day one assignments from column two of the table and so on for the remaining days of the second lesson and then last two observation lessons. Note, the percentages of lessons receiving two raters is not 25% in this abbreviated sample assignment table since we only present two days in the schedule of observations.

We chose the grouping of raters within rows so that the number of times pairs, triplets, and sets of four raters we assigned to the same row was roughly constant across all possible groupings. Hence, rater A and B were paired together in a row (i.e., would observe the same classroom) roughly the same number of times as rater A and C, rater A and D, rater B and C, and so on.

To assign classrooms to the rows of the assignment table, we created groups of classes in which travel time between classes was sufficiently short so that a SR could observe all the classrooms in a day even if they were in different schools. We then randomly assigned groups of classrooms to rows so that on a given day, the SRs assigned to the row would observe all the classrooms in the group. Groups contained one to three classrooms.

However, the plan was not fully implemented as designed. First, one SR was asked to leave the project. Second, because of the need to reschedule observations due to special events in schools and classes, the SR work plan did not always follow the schedule. Also, some of the SRs assigned for classes needed to be replaced in order to keep SRs busy and accommodate teacher's schedules. These re-assignments were ad hoc but attempted to retain the elements of the original

design: classes observed by three or four distinct SRs, one of whom observed two classes, one fifth of lessons double-scored, and the number of classes assigned a pairing, triplet or grouping of four SRs held roughly even across groupings. The fifth lesson also was assigned after the initial assignment of SRs also in an ad hoc manner designed to maintain the original design principals while balancing SR workloads.

We restricted the assignments of videos to SRs for scoring, so that a SR would not score a video from a lesson in a class that she observed. Using the structure of day blocks from our original observation assignment table we grouped together two rows from the assignment tables from the same day-block to create pairs of raters for each video assignment (see Table 2). The

Table 2

Example Rater Assignment Table for Video Scoring.

		Observation Period			
Classroom Group	Day	1	2	3	4
11	1	B	C	D	B
11	1	C	D	E	C
12	1	D	E	F	D
12	1	E	F	B	E
13	1	F	B	C	F
13	2	A	D	E	A
14	2	C	E	F	C
14	2	D	F	A	D
15	2	E	A	C	E
15	2	F	C	D	F

Note. The example shows that rows of Table A2 are combined to create pairs of ratings for each video. The video from lessons 1 for classroom group 11 will be assigned to raters B and C and the video from the second lesson will be assigned to raters C and D. The double coders from the live observations were dropped to create video assignments, and classroom group IDs are numbered 11 to 15 to indicate that raters will not be assigned videos from lessons for classroom groups they observed live.

video and live observations used the same assignments of rater labels from the tables to SRs. Because each SRs had distinct assignment on a day, pairing rows within a day resulted in pairs of two SRs per lesson. These groupings had good balance in terms of which SRs were grouped together to score the same video and to score videos from the same class. We randomly permuted the days to avoid ordering confounds between video and live scoring. We then randomly selected a set of possible assignments from this table for each class. If this randomly selected set of assignments had a SR scoring a video from an lesson in a class she observed, then we chose another row from the table. We made ad hoc changes to the initial assignments when one of SRs was removed from the study and for the fifth lesson. Again, the ad hoc changes maintained the principle feature of the scoring assignment design: all videos were assigned two researchers who had not observed the class lesson in person and we kept the pairings and grouping of researchers to the same class or video reasonably balanced so that groups occurred roughly the same number of times.

Table 3

Mean Differences between Mode (Live Score – Video Score) for all Dimensions and Domains.

Dimension/Domain Score	Mean Differences			
	Live - Video	SE	t	*p*
Positive Climate	0.357	0.037	9.53	< .001
Teacher Sensitivity	0.176	0.036	4.96	< .001
Regard for Adolescent Perspectives	0.327	0.038	8.67	< .001
Negative Climate	-0.013	0.023	-0.58	0.564
Behavioral Management	-0.143	0.033	-4.29	< .001
Productivity	-0.082	0.038	-2.15	0.032
Instructional Learning Formats	0.118	0.031	3.83	< .001
Content Understanding	0.184	0.037	5.03	< .001
Analysis and Problem Solving	0.444	0.040	11.10	< .001
Quality Feedback	0.335	0.049	6.88	< .001
Student Engagement	0.029	0.034	0.84	0.400
Emotional Support	0.287	0.031	9.35	< .001
Classroom Organization	-0.070	0.024	-2.96	0.003
Instructional Support	0.270	0.032	8.57	< .001

Table 4

Decomposition of Variability in Live Scores

	Classroom	Lesson	Segment	Rater	Rater x Classroom	Rater x Lesson	Residual
Positive Climate	29.4%	9.9%	6.9%	10.9%	8.1%	17.7%	17.1%
Teacher Sensitivity	19.9%	9.2%	10.4%	6.9%	0.0%	26.4%	27.3%
Regard for Adolescent Perspectives	10.4%	20.5%	11.8%	24.6%	0.0%	12.6%	20.0%
Negative Climate	23.1%	18.7%	10.8%	7.0%	4.0%	8.8%	27.7%
Behavioral Management	36.4%	11.1%	9.8%	8.3%	0.0%	20.0%	14.4%
Productivity	15.9%	2.0%	6.6%	17.1%	0.0%	34.1%	24.3%
Instructional Learning Formats	16.1%	8.4%	5.4%	11.8%	7.0%	19.3%	31.9%
Content Understanding	13.1%	13.3%	8.6%	16.2%	3.7%	16.9%	28.1%
Analysis and Problem Solving	4.4%	20.3%	5.7%	33.7%	6.0%	12.8%	17.1%
Quality Feedback	10.4%	8.5%	6.5%	42.1%	1.5%	13.6%	17.4%
Student Engagement	37.0%	6.2%	6.8%	3.8%	3.7%	20.5%	21.9%
Emotional Support	24.1%	16.6%	10.3%	15.9%	4.4%	16.2%	12.5%
Classroom Organization	35.0%	12.6%	9.5%	5.0%	0.0%	23.0%	15.0%
Instructional Support	13.4%	17.6%	6.5%	29.9%	7.6%	11.5%	13.5%

Table 5

Decomposition of Variability in Video Scores

	Classroom	Lesson	Segment	Rater	Rater x Classroom	Rater x Lesson	Residual
Positive Climate	20.8%	2.8%	4.9%	23.7%	4.2%	25.5%	18.1%
Teacher Sensitivity	18.1%	1.8%	8.1%	12.7%	3.9%	23.7%	31.8%
Regard for Adolescent Perspectives	10.8%	8.4%	10.2%	25.4%	1.5%	17.5%	26.3%
Negative Climate	22.7%	6.8%	8.3%	8.0%	8.4%	16.6%	29.3%
Behavioral Management	31.2%	5.8%	6.9%	13.6%	4.9%	20.7%	17.0%
Productivity	13.9%	1.7%	3.3%	22.6%	2.9%	30.3%	25.4%
Instructional Learning Formats	17.0%	6.3%	5.2%	10.8%	3.9%	16.0%	40.9%
Content Understanding	10.1%	6.2%	9.9%	17.7%	0.0%	26.0%	30.2%
Analysis and Problem Solving	4.7%	4.3%	4.2%	33.2%	0.0%	28.1%	25.4%
Quality Feedback	11.7%	2.4%	9.3%	32.2%	1.5%	21.5%	21.4%
Student Engagement	26.0%	2.9%	5.7%	9.9%	4.2%	21.6%	29.8%
Emotional Support	20.7%	4.0%	8.2%	23.1%	3.3%	23.5%	17.2%
Classroom Organization	29.4%	5.3%	7.9%	12.6%	6.5%	23.7%	14.6%
Instructional Support	13.2%	3.9%	9.1%	27.6%	1.9%	25.2%	19.3%

$$
\begin{aligned}
X_{clsr} = \quad & \mu && \text{[grand mean]} \\
+ \quad & \mu_c - \mu && \text{[classroom effect]} \\
+ \quad & \mu_l - \mu && \text{[lesson effect]} \\
+ \quad & \mu_s - \mu && \text{[segment effect]} \\
+ \quad & \mu_r - \mu && \text{[rater effect]} \\
+ \quad & \mu_{cl} - \mu_c - \mu_l + \mu && \text{[classroom by lesson]} \\
+ \quad & \mu_{cs} - \mu_c - \mu_s + \mu && \text{[classroom by segment]} \\
+ \quad & \mu_{cr} - \mu_c - \mu_r + \mu && \text{[classroom by rater]} \\
+ \quad & \mu_{ls} - \mu_l - \mu_s + \mu && \text{[lesson by segment]} \\
+ \quad & \mu_{lr} - \mu_l - \mu_r + \mu && \text{[lesson by rater]} \\
+ \quad & \mu_{sr} - \mu_s - \mu_r + \mu && \text{[segment by rater]} \\
+ \quad & \mu_{cls} - \mu_{cl} - \mu_{cs} - \mu_{ls} + \mu_c + \mu_l + \mu_s - \mu && \text{[classroom by lesson by segment]} \\
+ \quad & \mu_{clr} - \mu_{cl} - \mu_{cr} - \mu_{lr} + \mu_c + \mu_l + \mu_r - \mu && \text{[classroom by lesson by rater]} \\
+ \quad & \mu_{csr} - \mu_{cs} - \mu_{cr} - \mu_{sr} + \mu_c + \mu_s + \mu_r - \mu && \text{[classroom by segment by rater]} \\
+ \quad & \mu_{lsr} - \mu_{ls} - \mu_{lr} - \mu_{sr} + \mu_l + \mu_s + \mu_r - \mu && \text{[lesson by segment by rater]} \\
+ \quad & X_{clsr} - \mu_{cls} - \mu_{clr} - \mu_{csr} - \mu_{lsr} + \mu_{cl} + \mu_{cs}
\end{aligned}
$$

Figure 1. A G study model for decomposing the CLASS-S score X_{clsr} from a rating of one classroom (c) on one lesson (l), for one segment of the lesson (s) by one rater (r).