Validity and Reliability Information for the LSC Classroom Observation Protocol

The LSC Core Evaluation classroom observation protocol was developed to measure the quality of an observed science or mathematics classroom lesson. The protocol and the items it contains are based on standards of quality mathematics and science instruction as outlined in the *National Science Education Standards* (National Research Council, 1996) and the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989), *Professional Teaching Standards for School Mathematics* (NCTM, 1991), and *Assessment Standards for School Mathematics* (NCTM, 1995). Multiple reviews of the protocol by approximately 60 science and mathematics educators (who served as principal investigators of the Local Systemic Change projects) comprised the content validation of the protocol. Items identified by these individuals as not appropriately measuring the intended objective were revised and returned for further review. This process was iterated several times during the development of the instrument in order to assure broad agreement with the content of individual items and the integrity and completeness of the overall instrument.

The ratings section of the classroom observation protocol consists of a capsule description and several item sets measuring key features of the lesson. The capsule rating measures the overall quality of the observed lesson. Responses to the capsule rating are given on a seven-point scale from “Exemplary instruction” to “Ineffective instruction.” Four item sets measure key features of the lesson including lesson design (10 items plus one synthesis rating), lesson implementation (8 items plus one synthesis rating), mathematics/science content (9 items plus one synthesis rating), and classroom culture (9 items plus one synthesis rating). Responses for the individual items in these item sets are given on a five-point scale (1 = Not at all, 5 = To a great extent), with additional response options for “Don’t know” and “Not applicable.” Synthesis ratings are given on a five-point scale from “Not at all reflective of best practice in mathematics/science education” to “Extremely reflective of best practice in mathematics/science education.” An additional item set measuring likely impact of instruction on students’ understanding of mathematics/science includes 6 items with a five-point response scale from “Negative effect” to “Positive effect,” with additional response options for “Don’t know” and “Not applicable.” A separate section of the classroom observation protocol affords descriptive information about the classroom lesson as context for the ratings.

Reliability of raters using the classroom observation protocol is a fundamental concern. One of the major challenges facing observers when judging a lesson is assessing the effect of its placement in the instructional sequence. For example, when viewing a videotaped lesson where students were investigating questions of their choosing with inadequate controls, observers had different interpretations of the quality. Some saw this lesson as exemplary, assuming the teacher would use the inconsistent results as a springboard for discussion the next day, and then to repeating the experiments more carefully. Others considered the lesson a waste of time, and worse, worried that calling this type of activity science would lead to misconceptions about the nature of the scientific enterprise. While it is, of course, possible to interview teachers about how a single observed lesson fits into the sequence of instruction, or to observe a long enough sequence to judge for oneself, it is often not practical to do so, and certainly not for large numbers of teachers.

Thus, as part of the core evaluation, before using the protocol, prospective observers are required to participate in a two-day training session that consists of observing videotapes of classroom lessons and rating videotaped lessons. The purpose of the training is for prospective LSC observers to gain understanding of the standard 'rating keys' developed by a norming group of science and mathematics educators and to learn to rate lessons in accordance with the rating keys. In order to assess the reliability of trained raters, the ratings results of approximately 45 trained observers were compared with the norming group's rating key on a common observation of a videotaped lesson. On the seven-point capsule description scale, which reflects an overall judgment of the quality of mathematics or science instruction in the observed lesson, ninety-two percent rated the lesson within one rating level, either higher or lower, of the rating key standard. Fifty-seven percent of the trained observers rated the lesson in exact agreement with the rating standard. Individual rater reliability is considered acceptable for trained raters if their ratings fall within one category of the rating key standard.

The five item sets measuring lesson design, lesson implementation, mathematics/science content, classroom culture, and likely impact of instruction on students' understanding of mathematics/science were subjected to an internal consistency analysis. The internal consistency of these item sets is quite high. Using the full sample of classroom observations conducted for the LSC Core Evaluation in 1997-98, Cronbach's alpha coefficients for the item sets were as follows:

Table 1
Internal Consistency for Classroom Observation Protocol Item Sets with and without Synthesis Ratings

Item Set	Number of items	Cronbach's α without synthesis rating	Cronbach's α with synthesis rating	n (N = 625)
Lesson design	10 + 1	0.96	0.97	85
Lesson design, alternative[a]	9 + 1	0.92	0.94	358
Lesson implementation	8 + 1	0.94	0.95	308
Mathematics/science content	9 + 1	0.93	0.94	253
Classroom culture	8 + 1	0.93	0.94	109
Classroom culture, alternative[b]	7 + 1	0.92	0.94	515
Likely impact on students' understanding	6	0.94	—	521

[a] Fewer than half of observers responded to the item "Formal assessments of students were consistent with investigative mathematics/science" on the five point scale, as it is not applicable to many classroom observations. Due to the correspondingly low number of complete item set responses, this item was removed from the set in order to create an alternative item set with a larger number of complete responses.

[b] Fewer than half of observers responded to the item "Opportunities were taken to recognize and challenge stereotypes and biases that became evident during the lesson" on the five point scale, as it is not applicable to many classroom observations. Due to the correspondingly low number of complete item set responses, this item was removed from the set in order to create an alternative item set with a larger number of complete responses.