Measuring Classroom Assessment with a Work Sample

Andrea Beesley
McREL (Mid-Continent Research for Education and Learning)/REL-Central
4601 DTC Blvd., Ste 500
Denver, CO 80237
abeesley@mcrel.org
303-632-5541
April 15, 2009

Black and Wiliam (1998) conducted a review of quantitative studies in the area of classroom assessment and learning and determined that “innovations that include strengthening the practice of formative assessment produce significant and often substantial learning gains” (p. 140). Black and Wiliam’s 1998 review identified three key features of classroom assessment associated with gains in student achievement: providing accurate information, providing high-quality feedback to students, and involving students in the assessment process.

To attain accurate information about student performance, teachers must apply sound classroom assessment practices. First, teachers need to be able to understand and identify the purpose of their assessments. Teachers also need to provide their students with clear learning targets, in language that students can easily understand, so students comprehend the goals. Teachers also need to understand the different types of learning targets so that appropriate assessment methods can be used to collect accurate information (Stiggins, Arter, Chappuis, & Chappuis, 2004). Finally, teachers should provide effective feedback. Feedback represents the information that the teacher provides to the student in order to help the student reach the learning goal. In order to be effective, it is thought that feedback should be *task-involved* or *standards-oriented* and provide information regarding the task, not the individual or any reference group of individuals (Black & Wiliam, 1998; Butler, 1988). Effective feedback also models for students how to self-assess and consider and use the results themselves.

REL-Central is conducting an experimental study intended to measure the effects of one program of teacher professional development in classroom assessment, *Classroom Assessment for Student Learning* (Stiggins et al., 2004), on student achievement and other student and teacher outcomes. Schools participating in the study (n = 67) were randomly assigned to either the intervention or control group. All Grade 4 and 5 teachers who provide direct instruction in math in each intervention school have formed a team to implement the CASL program. Teachers in the control schools are engaging in their regular professional development activities. The student outcomes are scores on the state test and on a motivation survey (engagement, perceived autonomy, self-efficacy), and the teacher outcomes are knowledge of classroom assessment, student involvement in classroom assessment, and the work sample (practice of classroom assessment).

Despite evidence of the potential positive impact of classroom assessment on student achievement, recent research has revealed that teachers often do not receive much, if any, training in classroom assessment or related topics as part of their teacher preparation experience. Many teachers lack the knowledge, skills, and abilities needed to provide effective classroom assessment (e.g., Plake, Impara, & Fager, 1993). Given that applying the principles of effective classroom assessment is associated with increased achievement, professional development intended to increase teacher knowledge and skill in classroom assessment provides promise for increasing student achievement.

The study began in fall 2007 with baseline data collection. Training in the intervention occurred during Year 1, the 2007–2008 academic year. Data collection will conclude at the end of Year 2, the 2008–2009 academic year.

This paper describes a teacher work sample, intended to provide an accurate measure of teacher practice of classroom assessment in elementary mathematics, and describes the approach to using a panel to identify anchor papers.

Methods

Sample

Sixty-two schools in a Mountain West state with a total of 317 fourth- and fifth-grade teachers are participating in the study. Of these, 115 in the treatment group and 151 in the control group sent in the work sample package at baseline data collection in fall 2007.

Instrument

This study examines the effects of professional development in classroom assessment on the soundness of classroom assessment practices, including clear communication of learning targets, appropriate match of learning goals and assessment criteria, and feedback to students that describes strengths and needs in relation to learning targets. Systematically collecting samples of graded student work provides an efficient way (as compared to classroom observations) to find out what is happening in classrooms when teachers engage in assessment. Samples of graded student work directly capture teacher thinking and classroom assessment practice as opposed to the snapshot provided by a classroom visit and observation for which there is no guarantee that assessment practices will even occur. Procedures for collecting and analyzing these assessments and typical assignments were adapted from an artifact-based instrument developed to characterize classroom practice. This instrument, developed by researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), consists of an elementary and secondary language arts assignment rating system (Matsumura, Patthey-Chavez, Valdés, & Garnier, 2002).

Research on the data collection system suggests the data provide a reliable and valid characterization of classroom practice, at least as reliable as classroom observations. With clear scoring rubrics and training, inter-rater agreements have been acceptable (Borko, Stecher, Alonzo, Moncure, & McClam, 2005; Clare, Valdés, Pascal, & Steinberg, 2001; Matsumura, Garnier, Pascal, & Valdés, 2002).

In the work sample, teachers were asked to copy and send in three mathematics assignments that reflect their lesson objectives, with four examples of graded student work (including teacher feedback) for each assignment. The assignments included (1) a typical homework or seat-work assignment, (2) a typical in-class project or performance assessment, and (3) a typical quiz or end-of-week assessment. One of the homework or in-class assignments must ask students to show their work and explain their answers. Instructions to the assessment work sample asked teachers to attach the activity's directions and indicate via checklist and short-answer responses:

- The assignment and its learning goals.
- How the assignment fits with its unit.
- How it addressed the range of student skills with the assignment.
- How much time the students needed to do the assignment.
- The type of help the students received.
- How the assignment was assessed including scoring rubric.
- How the students performed on the assignment.

The study team will then use a rubric based on the work of CRESST (Matsumura, Patthey-Chavez et al., 2002), with two feedback dimensions added, to score the work samples on classroom assessment practices and quality of feedback to students. Qualified raters will be recruited using established procedures and criteria such as having combined experience and background in both classroom teaching and educational evaluation (Matsumura, Garnier, Pascal & Valdes, 2002). In order to maintain reliability and avoid any potential rater effects, raters will need to qualify prior to scoring work samples from the research study sample. Raters will need to achieve 80% exact agreement with the scores of the qualifying papers to qualify. The qualifying papers will include a set of work samples obtained from sample teachers that have known scores. When rating the research samples an inter-rater agreement of 80% for within one scale point is expected, given prior reports of artifact-based instruments (Matsumura, Patthey-Chavez et al., 2002).

The scoring rubric has 6 areas (focus of goals on student learning, clarity of grading criteria, alignment of learning goals and assignment, alignment of learning goals and grading criteria, type of feedback, and student involvement level of feedback). Teachers receive a score from 1 to 4 in each of the six areas based on all their submitted materials. Scores from the six areas will be combined, giving each teacher a single score from the work sample. Teacher work samples will be collected twice over the course of the study, at baseline and then again in spring 2009.

Scoring Panel Procedure

In order to identify anchor and qualifying papers, the researchers assembled a five-person panel: two retired professors who were assessment experts, two district-level personnel experienced in teaching and assessment, and one mathematics specialist. The panel convened in a one-day meeting in July 2008 to review the rubric, score some papers together as a panel, and then score other papers to be used as training and qualifying papers. Prior to the arrival of the panel the study team had assembled a set of anchor paper candidates thought to adequately represent all dimensions and levels of the rubric, so that the panel would not have to confront the entire sample and risk missing some of the higher-scoring papers, which were expected to be fairly rare in this baseline administration. When reviewing the rubric, the panel recommended slight changes to the wording in order to clarify the dimensions so that they could be used to score the papers. They then scored, as a whole group, one example of each type of assignment (homework/seatwork, end of week quiz or assessment, performance task or in-class project). Afterward they scored seven papers individually, with each panelist scoring each paper. The panelists decided to immediately accept scores agreed upon by four or five scorers, and confer among themselves to decide on those with less initial agreement. Afterward, the panelists scored 14 more papers by giving each paper three ratings (a paper was considered scored when any three panelists had scored it). They immediately accepted scores that were agreed upon by two out of three, and negotiated the others. The scoring of each paper was recorded by a facilitator, who transcribed each initial score from each panelist along with the negotiated score for each rubric dimension for each paper.

Outcomes of Scoring Panel

The panelists made changes to the rubric considered to be necessary to score the papers. For example, the dimension “clarity of grading criteria” was changed to make it necessary for the grading criteria to be clear to the students, not just the teachers. (At the presentation I will make available both versions of the rubric.)

The papers scored by all panelists, requiring four out of five identical scores for immediate agreement, showed a relatively low level of immediate agreement; the feedback dimensions had the highest levels of immediate agreement (see Table 1). The papers scored by three out of five panelists, requiring two out of three identical scores for immediate agreement (a less stringent criterion), showed greater immediate agreement. Overall the mean of the scores was 2.15 (SD = .61), indicating a fairly low level of ratings on the four-point rubric in this baseline sample. The lowest-scoring dimensions were the two feedback dimensions, with means of 1.67 and 1.62.

Table 1. Percentages of immediate agreement with 5 and 3 scorers.

Dimension	% immediate agreement (5 scorers), n=7	% immediate agreement (3 scorers), n=14
Focus of the goals on student learning	29%	86%
Clarity of the assessment criteria for students	14%	79%
Alignment of learning goals and task	14%	93%
Alignment of learning goals and assessment criteria	29%	71%
Feedback–type	43%	93%
Feedback–student involvement	43%	86%

Three papers were labeled by panelists as challenge papers, due to qualities that caused initial agreement of scores across the dimensions to be low. Those papers could be used in training as examples of papers whose relationships to the rubric dimension scoring levels are more ambiguous.

After the scoring panel completed their work, papers had been identified for most dimensions and levels of the rubric. However, level 4 papers were identified only for alignment of learning goals and task and alignment of learning goals and assessment criteria, and no level 1 papers were found for alignment of learning goals and task. The panelists had especial difficulty in finding papers that described the learning targets in language that was clear, explicit, and elaborated, and in finding papers that gave high-quality feedback, especially feedback that elicited student involvement.

Importance of the Study

While high-quality formative assessment has been shown to increase student achievement, it is difficult to measure teachers' actual practice of formative assessment in the classroom. The teacher work sample is intended to provide a window into classroom practice of formative assessment, while allowing for more information-gathering and greater efficiency as compared to classroom observations. This study describes the process of conducting the work sample with a relatively large number of teachers, and of conducting a scoring panel meeting to establish anchor papers (the results of the scoring of the rest of the work sample will be available at the time of the AERA conference). Researchers interested in measuring classroom assessment practices, along with school personnel who may want to use work samples as part of teacher assessment or professional development, can benefit from the resulting knowledge; they can learn what the process is and what to expect, and can compare their results to those in this study.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment, 10*(2), 73-104.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1-14.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (No. 545). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Matsumura, L. C., Garnier, H. E., Pascal, J., & Valdés, R. (2002). *Measuring instructional quality in accountability systems: Classroom assignments and student achievement*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Matsumura, L. C., Patthey-Chavez, G. G., Valdés, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103*(1), 3-25.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12, 39.

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right - using it well*. Portland, OR: Assessment Training Institute.