Measuring Instructional Quality in Accountability Systems: Classroom Assignments and Student Achievement

CSE Technical Report 582

Lindsay Clare Matsumura, Helen E. Garnier, and Jenny Pascal
CRESST / University of California, Los Angeles

Rosa Valdés
Los Angeles Unified School District

November 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.3 Indicators of Classroom Practice and Alignment
Joan Herman and Lindsay Clare Matsumura, Project Directors, CRESST/UCLA

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

MEASURING INSTRUCTIONAL QUALITY IN ACCOUNTABILITY SYSTEMS: CLASSROOM ASSIGNMENT QUALITY AND STUDENT ACHIEVEMENT

Lindsay Clare Matsumura, Helen E. Garnier, and Jenny Pascal
CRESST/University of California, Los Angeles

Rosa Valdés
Los Angeles Unified School District

Abstract

This report describes the technical quality of a CRESST-developed measure of the quality of classroom assignments piloted in the LAUSD's proposed new accountability system. For this study, 181 teachers were sampled from 35 schools selected at random. Participating teachers submitted three language arts assignments with samples of student work (N = 50). Results indicated a fair level of agreement among the raters who scored the assignments and a high level of internal consistency within four dimensions of assignment quality. The stability of the ratings and the number of assignments needed to yield a consistent estimate of quality differed by elementary and secondary school levels. As a group, secondary students who received higher quality assignments produced higher quality written work and scored higher on the reading and language portions of the Stanford Achievement Test, Ninth Edition, adjusted for student background and prior achievement.

Over the past two decades, numerous reform efforts and policies, ranging from professional development activities for teachers to the adoption of content standards for learning and instruction, have been implemented in public schools in efforts to improve instruction and student learning. Despite the range of reform programs in place intended to improve the quality of teaching, the success of these ventures generally has been assessed in one way—through student outcome scores on standardized tests of achievement. Less emphasis has been placed on monitoring the quality of instruction, the most important *school* factor influencing student achievement (Darling-Hammond, 2000; Tharp & Gallimore, 1988). Thus, educators and policymakers lack information on how instruction has (or has not) been influenced by reform efforts and how changes in specific aspects of classroom practice may (or may not) influence student learning.

The quality of instruction has not been directly measured in many accountability systems because few assessment tools exist that have the potential to measure the quality of classroom practice on a large-scale basis. Classroom observations are the most direct way to assess the quality of instruction, but these are time-consuming and expensive to conduct. Surveys are limited by the fact that they rely on teachers' (at times inaccurate) self-reports of their practice (Mayer, 1999; Spillane & Zeuli, 1999). Analyses of student work can provide useful information about student learning but do not necessarily provide information about students' opportunity to produce high-quality work in classrooms.

New indicators that help schools, districts, and states monitor and support efforts to improve the quality of instruction are clearly needed. These indicators are important for providing feedback to schools and districts about their interim progress toward reform goals. This is especially important, given the fact that numerous studies have shown that even when teachers "buy in" to a change in instructional practice, the classroom implementation of such a practice does not always reflect a reform program's intentions. This is true for efforts as diverse as teaching mathematics from a more conceptual perspective, implementing the process approach to writing instruction, and adopting content and performance standards for learning and instruction (Applebee, 1984; Briars & Resnick, 2000; Cohen & Ball, 1994; Matsumura, Patthey-Chavez, Valdés, & Garnier, in press; Spillane & Zeuli, 1999).

Indicators of classroom practice also are needed that draw attention to features of classroom practice that are germane to student learning (Linn & Baker, 1998). This is of critical importance in helping districts and schools choose how they might want to focus their professional development resources. Specifically, this is important in terms of providing information to schools and districts about specific areas of strength and weakness in classroom practice and what changes in instruction may have the greatest impact on student achievement.

For the past 4 years, CRESST has been developing indicators of classroom practice that potentially serve these purposes (Aschbacher, 1999; Clare, 2000; Clare, Valdés, Steinberg, & Pascal, 2001). The CRESST methodology is unique and features the collection of teachers' language arts assignments and associated student work and the application of a standardized rubric for measuring the quality of the assignments. The results of this scoring process then are used to create indicators of classroom practice.

To date, CRESST's research has focused on investigating the technical quality of this method in the context of a large-scale school reform initiative and in a small number of schools and classrooms (Clare & Aschbacher, 2001). In this report, the technical quality of the CRESST assignment measure is investigated with a larger number of schools and classrooms in the context of the Los Angeles Unified School District's (LAUSD) proposed accountability system piloted in a subsample of schools.

Background to the LAUSD's Local District Performance Measures Project

In July 2000, the LAUSD, the second largest school district in the country, divided into 11 "local districts," each with its own local district superintendent. To monitor the performance of the new local district superintendents, the LAUSD general superintendent commissioned the development of a new accountability system. This project, termed the Local District Performance Measures (LDPM), was developed by the Program Evaluation and Research Branch of the LAUSD (Cantrell et al., 2001).

Building on earlier systems of accountability undertaken by the district, the LDPM system was unique in that the intention was to measure both direct outcomes of student performance and the school processes expected to increase student performance (Cantrell et al., 2001). Specifically, the indicators used to measure student performance were the percentage of schools meeting the Academic Performance Index growth target and the percentage of schools meeting their expected matched student reading gain on the Stanford Achievement Test, Ninth Edition (Stanford 9). The indicators used to measure the school processes expected to increase student achievement included the percentage of schools reaching a satisfactory rating on the School Organization Index[1] and the percentage of schools reaching a satisfactory rating on the CRESST classroom assignment measure.

This report describes the technical quality of the CRESST classroom assignment measure used in the new LAUSD accountability system. Specifically, the report presents findings focusing on the reliability and stability of the assignment ratings, and the relationship of classroom assignment quality to student performance. The research questions addressed in this study are as follows:

[1] The School Organization Index is a survey that focuses on features of school organization expected to be associated with student achievement.

1. How reliable are the classroom assignment ratings?

2. How many assignments and raters are needed to obtain a consistent estimate of the quality of classroom practice?

3. What is the relation of the classroom assignment ratings to student performance (the quality of students' written work and Stanford 9 scores)?

Methods

Sample

Eligible teachers were recruited from Grades 4, 7, and 10 to participate in the district's initial pilot of the classroom assignment indicators during the 2000-2001 academic year (N = 181). These teachers were recruited from 35 schools that had been randomly chosen by the LAUSD across its 11 local districts. Of the teachers who were recruited, 50 returned assignments (26 elementary and 24 secondary teachers), a return rate of approximately 28%. Teachers had been teaching in the LAUSD an average of 7.86 years with a range of 1 to 32 years (n = 44; data were missing for 6 teachers). Of the teachers for whom we have data, the majority (88%) were female (n = 44). The ethnicity of these teachers was reported by the LAUSD as 59% White, 20% Latino, 14% African American, 5% Asian, and 2% Filipino.

Design, data collection, and rating procedures were based on prior CRESST research (Aschbacher, 1999; Clare, 2000; Clare et al., 2001). Raters (n = 6) were recruited by CRESST to score the assignments. All but one of these raters had classroom teaching experience. Expert raters (n = 2) who had been part of CRESST's classroom assignment project in previous years also scored assignments.

Student background information and reading and language achievement scores for years 2000 and 2001 were provided by the LAUSD (N = 3,668 students). Data were included in this study only for students from Grades 4, 7, and 10 in the year teacher assignment data were collected. Data for students identified with a special education code, however, were not included in this study. The final sample used in analyses with both teacher and student data consisted of 49 teachers (26 elementary, 23 secondary) and 2,577 students (614 elementary, 1,963 secondary). Fifty-two percent of the students were female (54% elementary, 52% secondary); 26% were designated as Limited English Proficient (43% elementary, 21% secondary); and 79% participated in a free lunch program during the year the teacher assignments were collected. The ethnicity of the students was reported by

the LAUSD as 7% White (13% elementary, 5% secondary); 70% Latino (66% elementary, 71% secondary); 13% African American (13% elementary, 13% secondary); 2% Asian (2% elementary, 2% secondary); 4% Filipino (2% elementary, 5% secondary); and 2% other or multiple ethnic identification (1% elementary, 2% secondary).

Procedures and Measures

Classroom assignments. LAUSD research staff contacted teachers by mail in November 2000. Each teacher received a packet of materials that included (a) a letter describing the purpose of the project, (b) assignment cover sheets to be completed by the teachers, (c) a copy of the rubric used to assess the quality of the classroom assignments, and (d) a short survey eliciting teachers' reactions to the data collection. Teachers were given 3 weeks to return the completed assignment materials and student work. All of the participating teachers were mailed follow-up letters a few weeks later. The first follow-up letter informed teachers that they would receive $75 for classroom instructional materials as compensation for the time they spent gathering the student work and completing the assignment materials. The second follow-up letter extended the deadline for sending in the materials by an additional 3 weeks. Randomly selected teachers also were contacted by phone and/or were visited at their schools (or received no support), in order to investigate which type of additional follow-up treatment would yield a higher response rate (see Cantrell et al., 2001, for more details about this portion of the data collection).

For each assignment, teachers were asked to complete a 2-page information sheet and submit four samples of student work—two considered to be of medium quality by the teacher and two considered to be of high quality by the teacher, for the class. A total of 139 assignments were collected from the 50 teachers who participated in the pilot study. One elementary school teacher and two secondary school teachers were missing one of the three assignments, and four secondary school teachers were missing two assignments. With the exception of the one secondary teacher for whom two assignments and all student data were missing, missing data were estimated for assignment quality ratings using mean substitution within grade level.

The criteria used for describing assignment quality were based on research that focused on instructional effectiveness (Porter & Brophy, 1988; Resnick, 1995; Slavin & Madden, 1989). CRESST also drew on standards for learning and instruction

(California Department of Education, 1998; 1999; Danielson, 1996), as well as the work of other researchers who have examined assignment quality (Newmann, Bryk, & Nagaoka, 2001; Newmann, Lopez, & Bryk, 1998; Peterson, 2001). Based on a review of this research, the following six dimensions were derived, each of which was rated on a 4-point scale (1 = *poor* to 4 = *excellent*).

Cognitive challenge of the task. This dimension describes the level of thinking required of students to complete the task. Specifically this dimension describes the degree to which students have the opportunity to apply higher order reasoning, engage with academic content material, and produce extended responses.

Clarity of the learning goals. This dimension describes how clearly a teacher articulates the specific skills, concepts, or content knowledge students are to gain from completing the assignment. The primary purpose of this dimension is to describe the degree to which an assignment could be considered a purposeful, goal-driven activity focused on student learning.

Clarity of the grading criteria. The purpose for this dimension is to assess the quality of the grading criteria for the assignment in terms of their specificity and potential for helping students improve their performance. How clearly each aspect of the grading criteria is defined is considered in the rating, as well as how much detail is provided for each of the criteria.

Alignment of the learning goals and task. This dimension focuses on the degree to which a teacher's stated learning goals are reflected in the design of the assignment tasks students are asked to complete. Specifically, this dimension attempts to capture how well the assignment appears to promote the achievement of the teacher's goals for student learning.

Alignment of goals and grading criteria. This dimension is intended to describe the degree to which a teacher's grading criteria support the learning goals, that is, the degree to which a teacher assesses students on the skills and concepts they are intended to learn through the completion of the assignment.

Overall quality. This dimension is intended to provide a holistic rating of the quality of the assignment based on its level of cognitive challenge, the clarity of the learning goals, the clarity of the grading criteria, the alignment of the learning goals and the assignment task, and the alignment of the learning goals and the grading criteria.

Novice raters participated in a 3-day training and scoring session at CRESST (1.5 days of training, and 1.5 days of scoring). The training involved reviewing the rubric and anchor assignments for each dimension and then scoring assignments from nonsample teachers. Raters were trained to assess the quality of assignments on six dimensions: cognitive challenge, clarity of the learning goals, clarity of the grading criteria, the alignment of goals and task, the alignment of goals and grading criteria, and the overall quality of the assignment. In order to streamline the scoring process, however, the two alignment dimensions were not scored, though they were included in the rating of overall quality.

Ratings were conducted separately for elementary and secondary school assignments, with three novice and two expert raters participating in each group. All assignments were scored by each of the raters within that level of schooling. Rater agreement was checked throughout the scoring period, and discrepancies in scores were discussed as appropriate. Raters took between 6 to 7 minutes on average to score each assignment on four dimensions. Other information about the classroom assignments (e.g., the length of time students took to complete the assignment, and the origin of rubrics teachers used to grade the assignments) also was recorded by CRESST and LAUSD research staff.

The quality of the assignments collected in the LAUSD ranged from poor to excellent at both the elementary (Grade 4) and secondary (Grades 7 and 10) levels. Most of the assignments, however, were considered to be of moderate quality (scored a 2) across the different dimensions (see Table 1).

Student work. Student work from the writing assignments was rated using three standards-based scales measuring organization, content, and MUGS

Table 1

Quality of Classroom Assignments (N = 50 Teachers)

	Elementary (n = 26)	Secondary (n = 24)
Quality of assignment ratings	M (SD)	M (SD)
Cognitive challenge of the lesson activities	2.05 (0.38)	2.24 (0.55)
Clarity of the learning goals	2.32 (0.43)	2.23 (0.43)
Clarity of grading criteria	2.10 (0.53)	1.88 (0.50)
Overall quality of the observed lesson	2.00 (0.37)	2.11 (0.53)

Note. Items were scored on a 4-point scale (1 = *poor*, 4 = *excellent*). Ratings were averaged across raters within each dimension of assignment quality.

(mechanics, use of language, grammar, and spelling). These scales were part of the Language Arts Project rubric developed by LAUSD and United Teachers-Los Angeles in partnership with CRESST at UCLA (Higuchi, 1996). Each of these dimensions was rated on a 4-point scale (1 = *poor* to 4 = *excellent*). Two researchers rated either the elementary or the secondary student writing assignments. Reliability was assessed by double-scoring at least 20% of the other corpus. Exact scale-point agreement for the elementary school set of students was 84% for the content scale, 84% for the organization scale, and 91% for the scale measuring the quality of writing mechanics (MUGS). Agreement for these scales at the secondary level was 88% for the content scale, 94% for the organization scale, and 88% for the MUGS scale.

Student achievement. Stanford 9 scaled scores were provided by the LAUSD for two consecutive years, 2000 and 2001, for the students of 49 teachers (data for one of the secondary teachers were not provided to us). Students' reading and language achievement scale scores used in the analyses were matched for years 2000 and 2001 with no missing data estimated for reading and language scores. For the hierarchical analyses used in this study, pair-wise missing data procedures were used when students were missing achievement scores or background information. Variable effects and variance components were based on all the data. The final sample used in the analyses with both teacher- and student-level data consisted of 49 teachers (26 elementary, 23 secondary) and 2,577 students (614 elementary, 1,963 secondary).

Analyses

Descriptive statistics were used to characterize the teachers' assignments and the quality of students' work. Cohen's kappa coefficients were calculated to investigate the proportion of agreement between raters after adjusting for chance agreement. Cronbach's alpha coefficients were calculated to estimate the internal consistency of the ratings (Abedi, 1996).

Generalizability studies were conducted to investigate the consistency of our classroom assignment ratings within dimensions and to investigate the design of the research study. Decision studies were conducted to explore alternative designs for future studies. Correlations were computed to measure both the strength of agreement between raters and the relationship of classroom assignment ratings to ratings of students' written work.

Finally, regression analyses were used to investigate the relationship between classroom assignment ratings and student reading and language achievement. Because teacher assignments were given to students in intact classrooms and students were nested within classrooms with reading and writing comprehension assignments given at the classroom level, a hierarchical design was used to predict student achievement. The effects of classroom assignment quality on students' reading and language scores on the Stanford 9 were estimated by using HLM version 5.0 (Raudenbush & Bryk, 2000). The three dimensions of assignment quality used to predict students' scores were the level of cognitive challenge, clarity of the teachers' learning goals, and the clarity of the grading criteria used to assess students' work. The scale measuring overall assignment quality was not included, since the analyses were aimed at estimating the separate effects of the three distinct dimensions of assignment quality. Analyses controlled for differences among classrooms in students' reading and language achievement scores from the prior year, and also for differences in gender, participation in a free lunch program, and language status (i.e., designation as Limited English Proficient).

A 2-level hierarchical analysis model was run separately on reading and language achievement. Two student-level (Level 1) models and one teacher-level (Level 2) model were estimated using maximum likelihood procedures. At Level 1, a random-effects model was analyzed to provide information on how much variation existed between and within students on reading and language achievement outcomes. The second Level 1 model adjusted reading and language achievement outcomes for individual student background. Student-level covariates were specified as predictors of reading and language achievement: reading (or language) scale scores from the previous year and dummy variables for students' participation in a free lunch program, language status, and gender. The covariates were centered around the grand mean for the year making them equal to the deviation from the grade-level mean.

At Level 2, the teacher-level model was expanded to include the three ratings of assignment quality (cognitive challenge, clarity of the learning goals, and clarity of the grading criteria) as predictors of the adjusted reading and language achievement means after taking into account differences in student background. Effects of assignment quality on student achievement and variance components were estimated.

Results

How Reliable Are the Classroom Assignment Ratings?

The percent of agreement between raters was calculated on assignment ratings within each grade level. Results indicated that there was a fair level of agreement among the five raters who scored the elementary assignments and among the five raters who scored the secondary assignments (see Tables 2 and 3). The percentage of time the five raters agreed on the exact scale point was low, but it was somewhat higher for the raters who scored the elementary assignments (26.9% to 57.7%) than for the raters who scored the secondary assignments (12.5% to 45.8%). This was especially true for the dimension assessing the overall quality of assignments at the secondary level (12.5% exact scale-point agreement across all three assignments). The percent agreement within one scale point, however, was very high for both groups (91.7% to 100%).

Table 2

Reliability of Rating Scales for the Elementary Classroom Assignments (*n* = 26 Teachers)

Scale	Kappa[a]	Alpha	% Exact agreement 5 raters	% Agreement within 1 scale point
Assignment type: Writing				
Cognitive challenge	.42	.87	38.5	92.3
Clarity of the learning goals	.53	.92	57.7	100
Clarity of grading criteria	.53	.94	38.5	88.5
Overall quality	.41	.89	38.5	96.2
Assignment type: Reading (#1)				
Cognitive challenge	.43	.89	34.6	100
Clarity of the learning goals	.44	.88	34.6	96.2
Clarity of grading criteria	.53	.94	42.3	88.5
Overall quality	.39	.85	34.6	96.2
Assignment type: Reading (#2)				
Cognitive challenge	.56	.94	38.5	100
Clarity of the learning goals	.51	.91	38.5	100
Clarity of grading criteria	.55	.96	42.3	96.2
Overall quality	.46	.92	26.9	100

[a] Kappa coefficients are significant at $p < .01$.

Table 3

Reliability of Rating Scales for the Secondary Classroom Assignments (n = 24 Teachers)

Scale	Kappa[a]	Alpha	% Exact agreement 5 raters	% Agreement within 1 scale point
Assignment type: Writing				
Cognitive challenge	.51	.93	41.7	100
Clarity of the learning goals	.50	.92	45.8	100
Clarity of grading criteria	.40	.91	25.0	100
Overall quality	.32	.92	12.5	100
Assignment type: Reading (#1)				
Cognitive challenge	.59	.95	41.7	100
Clarity of the learning goals	.50	.93	37.5	95.8
Clarity of grading criteria	.37	.90	20.8	95.8
Overall quality	.36	.93	12.5	100
Assignment type: Reading (#2)				
Cognitive challenge	.43	.92	29.0	91.7
Clarity of the learning goals	.50	.92	33.3	95.8
Clarity of grading criteria	.43	.94	29.2	100
Overall quality	.35	.93	12.5	100

[a] Kappa coefficients are significant at $p < .01$.

Kappa coefficients were calculated to investigate whether the pattern of agreement observed was greater than would be expected if the raters had randomly assigned scores. Significant kappas for each dimension for each assignment ($p < .01$ level or higher at both levels of schooling) indicated that the level of rater agreement was better than chance. The magnitude of the kappas ranged from .39 to .59, however, indicating only a fair level of agreement among the five raters (Fleiss, 1981). Alpha coefficients also were calculated to investigate the internal consistency of the ratings within each assignment for each dimension. This statistic considers the trend in rater agreement and ranged from .87 to .96, confirming a high level of internal consistency within each dimension for each assignment type.

Though the overall level of agreement among the five raters in both groups was acceptable (but only moderate) overall, the level of agreement between individual pairs of raters varied considerably. For example, the correlation between the novice raters and the expert raters on ratings of cognitive challenge ranged from

.92 to .41 for the elementary assignments and from .82 to .30 for secondary assignments. Not surprisingly, the two expert raters, who had participated in CRESST's previous research, had the highest level of agreement. The novice raters who had the highest level of agreement with the expert raters had some experience as classroom teachers combined with some background in educational evaluation. The raters with only classroom teaching experience, in contrast, had the lowest level of agreement with the expert raters.

How Many Assignments and Raters Are Needed to Obtain a Consistent Estimate of the Quality of Classroom Practice?

To answer this question, the consistency (or stability) of the ratings for each dimension across the different assignment types was estimated using our design with three teacher assignments and five raters (Abedi, 1996). As illustrated in Table 4, the consistency of the ratings across assignments collected in elementary schools ranged from fair (.65 for the clarity of learning goals) to poor (.45 for cognitive challenge). The ratings of the secondary assignments, in contrast, yielded more consistent estimates of quality and ranged from .78 (clarity of the grading criteria) to .89 (cognitive challenge and overall quality).[2]

Generalizability and decision studies also were conducted using the GENOVA program to determine how many raters and assignments would be necessary to obtain a stable estimate of the quality of classroom practice (see Tables 5 and 6). In the generalizability study (G study), the design of five raters and three teacher assignments yielded a generalizability coefficient (G coefficient) of only .46 for the elementary school assignments (.80 and above is considered to be good), which replicated the pattern of results for the individual dimensions. The low G coefficient could be related to the high degree of variation *within* elementary teachers in the quality of the assignments. As illustrated in Table 5, 44.3% of the total variance was explained by the interaction of teacher by assignment type, far eclipsing the variation between teachers (15.9% of the total variance explained). In other words, individual teachers at the elementary school level tended to submit assignments of differing quality.

[2] We questioned whether the disparity in the results for the different levels of schooling was a result of our having imputed data for more assignments collected at the secondary level than at the elementary level. The analyses were rerun without the teachers for whom we imputed data for two assignments (n = 4 teachers). Analyses produced similar generalizability coefficients indicating that the pattern of results shown in Table 4 was due to factors other than imputed scores.

Table 4

Stability of Rating Scales for the Classroom Assignments (N = 50 Teachers)

Dimension	G coefficient
School level: Elementary (n = 26)	
Cognitive challenge	.45
Clarity of the learning goals	.65
Clarity of grading criteria	.62
Overall quality	.46
School level: Secondary (n = 24)	
Cognitive challenge	.89
Clarity of the learning goals	.84
Clarity of grading criteria	.78
Overall quality	.89

Table 5

Estimated Variance Components and Percent of Variance Explained by Teacher, Assignment Type, and Rater, Elementary Level (n = 26 Teachers)

	Variance component	% of variance explained
Teacher	.0696	15.9
Assignment Type	.0000	0.0
Rater	.0125	2.9
Teacher x Assignment Type	.1936	44.3
Teacher x Rater	.0365	8.4
Assignment Type x Rater	.0000	0.0
Teacher x Assignment Type x Rater	.1247	28.5

Table 6

Estimated Variance Components and Percent of Variance Explained by Teacher, Assignment Type, and Rater, Secondary Level (n = 24 Teachers)

	Variance component	% of variance explained
Teacher	.4327	57.0
Assignment Type	.0180	2.4
Rater	.0179	2.4
Teacher x Assignment Type	.0887	11.7
Teacher x Rater	.0664	8.7
Assignment Type x Rater	.0000	0.0
Teacher x Assignment Type x Rater	.1360	17.9

In contrast, our design yielded a higher G coefficient (.88) for the assignments collected at the secondary level. As shown in Table 6, most of the variation was found between teachers at that level of schooling (57% of the total variance) rather than in the interaction of teacher by assignment type (2.4%). This pattern is consistent with our results from previous years (Clare, 2000; Clare et al., 2001).

We next conducted decision studies in order to estimate G coefficients for varying numbers of assignments and raters. Given the likelihood that assignments collected in a large-scale study would not be scored by more than two raters, analyses focused on potential designs that utilized only one or two raters (see Table 7). Results revealed that a design of only two raters and three assignments at the *elementary* school level still might not yield a stable estimate of quality (G = .50). A design of only two raters and three assignments could, however, yield a stable estimate of quality at the *secondary* level (G = .84).

What Is the Relation of the Classroom Assignment Ratings to Student Performance?

Based on our results indicating that the classroom assignment ratings lacked stability at the elementary school level, the relation of classroom assignment ratings to measures of student achievement was investigated at the secondary level only.

Additionally, based on our results indicating a fair level of overall reliability for the classroom assignment rating scales and considerable variation between raters, the ratings of classroom assignments used in subsequent analyses were averaged across the three raters with the highest level of agreement (the two expert raters and one novice rater). In the next set of analyses, the quality of teachers' writing

Table 7

Estimated G Coefficients Based on the Number of Assignments and Raters (N = 50 Teachers)

Number of assignments	Number of raters	Estimated G coefficient (elementary)	Estimated G coefficient (secondary)
3	1	.33	.75
3	2	.40	.84
4	1	.37	.78
4	2	.46	.86
5	1	.41	.80
5	2	.50	.87

assignments was correlated with the quality of students' writing and then used to predict students' scores on the reading and language portions of the Stanford 9 (adjusted for students' prior achievement test scores, language status, and participation in a free lunch program).

Students' written work. Of the 24 secondary teachers who submitted writing assignments, 19 returned student work that could be scored for this analysis (e.g., assignments that required students to fill in blank spaces on a worksheet could not be scored using our writing rubric; see Table 8).

As illustrated in Table 9, the quality of the content of secondary students' work was associated with the level of cognitive challenge ($r = .76$, $p < .01$), the clarity of teachers' goals ($r = .50$, $p < .05$), and the overall quality of classroom assignments ($r = .72$, $p < .01$). The quality of the organization of students' work was associated with the level of cognitive challenge of the assignment ($r = .58$, $p < .01$), the clarity of the grading criteria ($r = .48$, $p < .05$), and the overall quality of the assignment ($r = .57$, $p < .05$). The quality of the mechanics of students' writing (MUGS) also was associated

Table 8

Quality of Secondary Students' Written Work (n = 19 Teachers)

	M (*SD*)	Range
Content	1.76 (.64)	1.00-3.00
Organization	1.73 (.63)	1.00-3.00
Writing mechanics (MUGS)	1.94 (.66)	1.00-4.00

Note. Items were scored on a 4-point scale (1 = *poor*, 4 = *excellent*). MUGS = mechanics, use of language, grammar, and spelling.

Table 9

Relationship of the Quality of Student Work and the Quality of Teachers' Assignments in Secondary School Classrooms (n = 19 Teachers)

	Classroom assignment variables			
Student work variables	Cognitive challenge	Clarity of the goals	Clarity of grading criteria	Overall quality
Content	.76**	.50*	.47	.72**
Organization	.58**	.42	.48*	.57*
Writing mechanics (MUGS)	.61**	.41	.47*	.60*

Note. Assignment ratings are based on the average scores of the raters.

* Significant at $p < .05$. ** Significant at $p < .01$.

with the level of cognitive challenge ($r = .61$, $p < .01$), the clarity of the grading criteria ($r = .47$, $p < .05$), and the overall quality of the assignment ($r = .60$, $p < .05$).

Stanford 9 (Reading). Variance components estimated in the first random effects student-level model indicated significant variability in students' reading achievement scores plus significant within-student variance. In the second student-level model, variance components indicated that adding covariates substantially reduced the within-student variance on reading achievement scores 56%. Both higher reading achievement scores in the prior year and English language proficiency were positively associated with higher adjusted reading achievement outcomes (see Table 10). However, after controlling for students' background covariates, significant variation among students' reading achievement scores still remained to be explained.

Analyses of the teacher-level model showed that the quality of teachers' assignments predicted 19% of the variance in adjusted reading achievement scores (see Table 11). Students' exposure to assignments that were more cognitively challenging and had less clear learning goals predicted higher adjusted reading achievement outcomes for students.

One possible explanation for the negative regression coefficient for the clarity of learning goals dimension may be found in the strong correlations among predictor variables, indicating that the data exhibited multicollinearity. The correlation between clarity of the learning goals and level of cognitive challenge (summed across the one writing and two reading assignments) was .74, and between clarity of the learning goals and clarity of grading criteria was .69. To further understand the association between the clarity of goals and reading achievement, we estimated the adjusted reading achievement scores for secondary students specifying clarity of goals as the sole predictor. Clarity of goals, on its own, did not significantly predict adjusted reading achievement, indicating that the relationship changed substantially when the three strongly correlated assignment quality ratings were included as joint predictors of reading achievement.

Stanford 9 (Language). Variance components estimated in the first random effects student-level model identified significant variability in students' language achievement Stanford 9 scores plus significant within-student variance. In the second student-level model, adding covariates substantially reduced the within-

Table 10

Results for Controlling the Effects of Student Gender, Participation in Free Lunch, and Language Status on Secondary Students' Reading and Language Scale Scores (N = 23 Teachers, 1,963 Students)

Covariates	Reading scale score: Year 2001	Language scale score: Year 2001
Reading or language scale score, Year 2000	0.72*** (0.03)	0.63*** (0.02)
Student gender	0.61 (0.81)	0.94 (0.85)
Participation in free lunch program	-0.13 (1.02)	-2.68* (1.20)
Language status (designation as Limited English Proficient)	-5.29*** (1.17)	-5.29*** (1.07)

* Significant at $p < .05$. ** Significant at $p < .01$. *** Significant at $p < .001$.

Table 11

Results for Estimating the Effect of Teacher Assignment Quality on Secondary Students' Reading and Language Scale Scores (N = 23 Teachers, 1,963 Students)

Predictors	Reading scale score: Outcome	Language scale score: Outcome
Intercept	663.69*** (1.05)	649.25*** (1.15)
Quality of assignment:		
Cognitive challenge	2.69** (0.86)	-0.56 (0.84)
Clarity of the learning goals	-3.01* (1.34)	-0.77 (1.30)
Clarity of grading criteria	0.70 (0.89)	1.89* (0.87)

* Significant at $p < .05$. ** Significant at $p < .01$. *** Significant at $p < .001$.

student variance on Stanford 9 language achievement scores 47%. Both language achievement scores in the prior year and English language proficiency were significantly associated with higher language achievement (see Table 10). After controlling for student background covariates, significant variation among student learning achievement scores still remained to be explained.

Analyses of the teacher-level model showed that quality of teachers' assignments predicted 8% of the variance in adjusted language achievement outcomes. Secondary students' exposure to teachers' assignments with higher clarity of grading criteria was positively related to higher adjusted language achievement outcomes (see Table 11).

Summary and Conclusions

In summary, the CRESST assignment measure appears to measure important aspects of classroom practice that are germane to student learning, at least at the secondary level. This study determined that classroom assignment ratings were reliable estimates of the quality of assignments. A fair level of agreement was found among the raters who scored assignments, along with a high level of internal consistency within each quality dimension for each assignment type. While the level of agreement was acceptable overall among the five raters in both the elementary and secondary school groups, the level of agreement between each pair of raters varied considerably. Not surprisingly, the two expert raters, who had been part of CRESST's previous research, had the highest level of agreement. The raters with the lowest level of agreement had only classroom teaching experience. Such variation in rater reliability raises important issues regarding rater training on a large scale. It may be necessary to pre-screen raters to assess reliability before scoring the sample corpus. Excluding raters with low levels of interrater agreement could improve overall reliability and stability of the assignment ratings. This may be especially important when large numbers of raters are needed to score assignments and the raters may lack evaluation experience or experience applying rubrics.

Results of G studies and decision studies indicated a mixed picture with regard to the stability of classroom assignment ratings at the different levels of schooling, and with regard to the number of assignments needed to yield a reliable and consistent estimate of quality. Specifically, our design in which teachers submitted three assignments that were assessed by five raters yielded a stable estimate of quality at the secondary school level, but not at the elementary school level. In contrast to the secondary teachers, there was more variation within elementary school teachers in terms of assignment quality than between teachers. The variation within elementary school teachers may be related to a pattern of these teachers submitting a mixture of commercially produced assignments and rubrics along with assignments and rubrics they created themselves. Specifically, 27% of the elementary school writing assignments and 59% of the reading comprehension assignments were generated from outside sources, including Open Court (2000, SRA / McGraw-Hill) worksheets (14%) and CRESST performance assignments (10%). Additionally, the majority of the writing assignments (58%) at the elementary level were assessed using rubrics generated by outside sources. These outside sources included the school district or the Los Angeles Annenberg Metropolitan Project School Family

(15%), teachers at the school (15%), and published instructional programs or teachers' guides (19%). Secondary teachers, in contrast, generated nearly all of their own assignments and scoring materials.

The findings presented here are limited, however, by the small sample size. Analyses based on a larger sample of teachers could yield more stable results. Considering the interaction of teacher and assignment type at the elementary school level, collecting only one type of assignment (for example, multiple writing assignments as opposed to a combination of assignments) also could yield a more consistent estimate of quality. More research is needed to explore this possibility, as well as the technical quality and utility of this method used when teachers submit commercially produced assignments.

The instability in the elementary-level assignment ratings restricted the analyses of the quality of teachers' assignments and student achievement to the secondary level. Although limited, the results indicated that higher quality teachers' assignments were associated with higher quality student work, a finding consistent with CRESST's previous research. Moreover, higher quality of teachers' assignments predicted higher student scores on the reading and language portions of the Stanford 9, even after controlling for students' backgrounds and prior level of achievement.

Contrary to expectations, the clarity of teachers' learning goals generally predicted lower adjusted reading achievement. This negative relationship was likely the result of multicollinearity, that is, strong associations among the predictor variables. The power of the clarity of goals dimension to positively predict adjusted reading achievement may have been reduced by the inclusion of the cognitive challenge dimension. Part of its positive predictive power was already captured by cognitive challenge. In other words, teachers who created more cognitively challenging assignments also were likely to have clearer goals, and both qualities were related to higher reading achievement. The variance in adjusted reading achievement not captured by the cognitive challenge dimension then might have been explained by teachers' assignments that did exhibit clear goals but were not very cognitively challenging. In other words, assignments that were not very cognitively challenging but for which the teacher had clear goals likely would be associated with lower reading achievement. Additional research is needed with larger samples of elementary and secondary teachers and their students to further investigate the relationship between the clarity of teachers' goals and students'

achievement. Adjustment of this dimension, for example, focusing on the content of teachers' goals in addition to their clarity, also may be necessary.

The two remaining dimensions measuring classroom assignment quality, however, were associated with improved student achievement. Specifically, students who were exposed to assignments that had clearer and more articulated grading criteria also received higher adjusted language achievement scores on the Stanford 9. And students scored higher on the reading portion of the Stanford 9 when they had been exposed to more cognitively challenging assignments. These findings suggest that the classroom assignment measure is sensitive to aspects of instructional practice that make a difference in student achievement. This could have implications for teachers' professional development, as well as for school personnel and policymakers who are seeking to implement measures of classroom practice in future accountability and evaluation efforts. More research is needed to confirm these results with larger samples of teachers and their students. If the measure is to be used for professional development purposes, additional research also is needed to investigate whether improving the quality of teachers' assignments alone would positively influence student learning, and to develop materials based on this method that would be useful and accessible to teachers.

References

Abedi, J. (1996). Interrater/test reliability system (ITRS). *Multivariate Behavioral Research, 31,* 409-417.

Applebee, A. N. (1984). *Contexts for learning to write: Studies of secondary school instruction.* Norwood, NJ: Ablex.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Tech. Rep. No. 513). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Briars, D. J., & Resnick, L. B. (2000). *Standards, assessments—and what else? The essential elements of standards-based school improvement* (CSE Tech. Rep. No. 528). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

California Department of Education. (1998). *California standards for the teaching profession* (Developed by the California Commission on Teaching Credentialing and the California Department of Education). Sacramento, CA: Author.

California Department of Education. (1999). *Reading/language arts framework for California public schools, kindergarten through grade twelve* (Developed by the California Development and Supplemental Materials Commission). Sacramento, CA: Author.

Cantrell, S., Lyon, N., Valdés, R., White, J., Recio, A., & Matsumura, L. C. (2001, June). *Pilot study report: The Local District Performance Measures* (Report submitted by the Program Evaluation and Research Branch of the Los Angeles Unified School District). Los Angeles: Los Angeles Unified School District.

Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice* (CSE Tech. Rep. No. 532). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7,* 39-59.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. (2001). *Teachers' assignments as indicators of classroom practice in elementary schools* (CSE Tech. Rep. No. 545). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cohen, D. K., & Ball D. L. (1994). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12,* 249-256.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2000, January). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives, 8*(1). Retrieved February 28, 2001, from http://olam.ed.asu.edu/epaa/v8n1

Fleiss, J. L. (1981). *Statistical methods for raters and proportions.* New York: Wiley.

Higuchi, C. (1996). *Improving student learning: High standards, standards-based curriculum, and standards-based assessment models.* Los Angeles: University of California, National Center for Research in Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L., & Baker, E. L. (1998, Fall). School quality: Some missing pieces. *CRESST Line,* pp. 1-3, 7.

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

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21,* 29-45.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago: Consortium on Chicago School Research.

Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report.* Chicago: Consortium on Chicago School Research.

Peterson, A. (2001). NAEP/NWP study shows link between assignments, better student writing. *The Voice: A Newsletter of the National Writing Project, 6*(2), 1, 16-17.

Porter, A. C., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the Institute for Research on Teaching. *Educational Leadership, 45*(8), 74-85.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks CA: Sage.

Resnick, L. B. (1995). From aptitude to effort: A new foundation for our schools. *Daedalus, 124*(4), 55-62.

Slavin, R., & Madden, N. (1989). What works for students at risk: A research synthesis. *Educational Leadership, 46*(5), 4-13.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21,* 1-27.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context.* New York: Cambridge University Press.