Evaluating the Effects of an Integrated Assessment System on Science Teachers' Assessment Perceptions and Practice

Lily Roberts and Mark Wilson

University of California, Berkeley

Mailing Address:

Graduate School of Education
3659 Tolman Hall
University of California
Berkeley, CA 94720-1670

Telephone:

Lily Roberts -- (510) 642-7968
Mark Wilson -- (510) 642-7966

Submitted for Review:

April 6, 1998

Running head: Evaluating the Effects of an Integrated Assessment System

Key words: mixed methods evaluation; multilevel, multisite evaluation; science assessment

Evaluating the Effects of an Integrated Assessment System on Science Teachers' Assessment Perceptions and Practice

Abstract

In this paper, we describe the use of a mixed method design to evaluate the impact of an integrated assessment system on science teachers' assessment perceptions and practice. This design was selected to address the methodological challenges of multilevel and multisite evaluation presented by a complex innovation. We provide an example of a cross-level exploratory analysis, facilitated by the integration of quantitative and qualitative methods; this provides one solution to the interpretation of multilevel, multisite evaluation data when a purposive sample is too small to use sophisticated quantitative techniques. The findings indicate that aggregated results would be misleading and discrepant cases would be masked.

Evaluating the Effects of an Integrated Assessment System on
Science Teachers' Assessment Perceptions and Practice

The Science Education for Public Understanding Program received funds from the National Science Foundation (NSF) to develop and field-test a year-long, middle school science course with an integrated assessment system, entitled Issues, Evidence and You (IEY). The SEPUP Assessment Project was designed to apply new theories and methodologies in the field of assessment to the practice of teacher-managed, classroom-based assessment of student performance (Wilson & Adams, 1996). In this paper, we describe the use of an integrated mixed method design to evaluate SEPUP's impact on science teachers' assessment practices. This design was selected to address the methodological challenges of multilevel and multisite evaluation presented by this complex innovation.

In congruence with the National Science Education Standards (National Research Council, 1996), the SEPUP IEY curriculum is designed to engage students in an "issues-oriented, hands-on" approach to thinking about scientific issues that are relevant to their daily lives (e.g., water, waste, energy, and environment) (SEPUP, 1996). Further, student understanding is assessed in an ongoing manner using the components of the SEPUP Assessment System (Roberts, Wilson, & Draney, 1997). The teachers' role is that of facilitator in the students' development. The aim is that assessment information becomes a scaffolding mechanism for instructional change that further facilitates student learning. Tobin (1995) indicates that any "critical approach" to teacher development must provide teachers with "autonomy to identify problems in their classrooms and seek solutions that make sense to them" (p. 147). With SEPUP assessment, the long-term goal is for teachers to become autonomous assessors of their students' understanding of science, which will be evidenced by their own professional development in terms of changing assessment practices. To work toward this ideal, we embedded the evaluation in the development and implementation of the SEPUP Assessment System.

In SEPUP, teachers engage in local assessment moderation, a scorer calibration process that also serves as a mechanism for teacher professional development (Ingvarson, 1990; Linn,

1994; Roberts, Sloane, & Wilson, 1996). Moderation provides a critical forum for teachers to function as a "community of judgment" (Wilson, 1994) while they learn about alternative assessments, engage in collegial discourse and reflect on their instructional practice. According to Harmon (1995), alternative assessments will be implemented "only if teachers understand their use and the depth of the content they demand, are empowered to make instructional decisions, and are supported by school districts which encourage teacher change" (p. 46). Consequently, moderation was added to ensure that the alternative science assessments would actually be used.

Evaluation Design: Integrated Mixed Methods

The evaluation of SEPUP presented a common methodological challenge encountered in program evaluation: How do you address the multilevel, multisite nature of complex innovations? As admonished by Burstein (1980), educational research and evaluation will not align with reality if we do not explicitly address the multilevel structure of our data. Further, Chen (1997) suggests that the choice of evaluation design "depends on the contextual circumstances" that surround the particular project to be evaluated (p. 63). For this study, an integrated mixed-method design was used because the framework for the evaluation is a multilevel, multisite model (i.e., teachers nested within SEPUP Centers across the U.S.). This type of mixed-methods design can be differentiated from a component design (e.g., the use of mixed methods to triangulate the results to achieve greater validity) because it achieves a greater integration of the different methods used (Caracelli & Greene, 1997). This study can be characterized as an embedded or nested mixed-methods design, which combines both qualitative and quantitative methods within a hierarchical model to explore the affect of this complex innovation on teachers' assessment practices.

Under ideal circumstances, a quantitative design using multilevel modeling would be appropriate. A Hierarchical Linear Model (HLM) statistical analysis, for example, can address both multilevel (teachers nested within Centers) and multisite (variation in Centers on the Center-level predictor variables) issues within the same analytical framework. HLM could be used to: (1) examine teacher change on the variables of interest; (2) compare mean differences on teacher

change between Centers; and (3) assess within-Center variation attributable to Center-level predictor variables. The first two analyses can also be performed using standard statistical approaches. However, if we were to use only quantitative methods, the third analysis would not be possible in the SEPUP case, because HLM or an equivalent statistical analysis could not be used due to insufficient sample size. Given the developmental nature of SEPUP, both the numbers of Centers (or groups) and the numbers of teachers within Centers is too sparse.

To address the third issue explicitly, we employed a mixed-methods design. The qualitative analysis provided evidence regarding Center-level variations. This information about the Center-level predictor variables was then integrated with the quantitative approach to provide a substantive interpretation of the relationship between teacher outcome variables and the contextual factors at the Center level. By combining the evidence in this way, we were able to conduct cross-level exploratory analyses as recommended by Bryk and Raudenbush (1993).

Subjects

Three treatment groups were identified for this evaluation: SEPUP Assessment Development Centers (ADCs); SEPUP Professional Development Centers (PDCs); and a non-SEPUP comparison group. All of these teachers taught science in grades 7, 8 or 9. All teachers had been nominated based on their excellence as science teachers and were willing to participate in the field test. The comparison teachers were the "next on the list" who did not become part of the field test. The distinction between ADC and PDC was at the Center-level rather than the teacher level.

All SEPUP teachers used the Issues, Evidence and You curriculum and the non-SEPUP teachers used their traditional science curriculum. The SEPUP ADC teachers received the most comprehensive treatment, both the curriculum and the assessment tools designed for SEPUP. The ADC teachers were supported in their use of the SEPUP Assessment System through the use of local assessment moderation meetings. The SEPUP PDC teachers received the curriculum containing the embedded assessments, but no support for using the various components of the assessment system.

The sample consists of all teachers involved in the 1994-95 field test of SEPUP's IEY. There were 26 SEPUP teachers and seven comparison teachers in six ADCs. In the six PDCs, there were 25 SEPUP teachers and five comparison teachers. Centers were organized a bit differently in each location. Some Centers were located in a single, large school district and a district representative functioned as the ADC director. Other Centers were organized less centrally, for example, one ADC had five SEPUP teachers from five different districts, and one of these teachers also served as the ADC director. In some cases, the Center director was not from the local school or district, but rather a university person involved in science education.

Evaluation Research Questions

The research questions correspond to the multilevel, multisite nature of the SEPUP evaluation. The first two questions focus on the individual level, examining the effect that program participation has had on teachers: Were the changes for the SEPUP teachers greater than those for the non-SEPUP teachers in terms of assessment perceptions and practice? Were the changes for the SEPUP ADC teachers greater than those for the SEPUP PDC teachers in terms of assessment perceptions and practice?

The last two questions focus on the group level, specifically looking at the teachers in the Assessment Development Centers (ADCs) who used the assessment system and participated in local assessment moderation. These latter two questions examine the organizational factors that are associated with teacher change: What features of the ADCs are associated with differences in teacher change among Centers? How do the ADCs differ in terms of teacher change?

Description of the Instrument

The SEPUP Inventory of Teachers' Assessment, Collegial, and Instructional Practices (SITACIP) was designed and validated for this evaluation study (Roberts, 1996). The SITACIP survey is a 77-item self-report of teachers' assessment, instructional and collegial practices as well as their perceptions about the usefulness of various assessment strategies. The subsections of the SITACIP fall into three basic question types: (a) frequency of use of various instructional and assessment strategies; (b) Likert-type scales for both collegiality and reasons for assessment

strategy choices; and (c) attitudes about the usefulness of different assessment strategies for assessing learning, guiding instruction, and grading.

This paper focuses on the assessment-related SITACIP scales: (1) teachers' perceptions of the usefulness of alternative assessment strategies for developing an understanding of what students know about science (Assessing Learning); (2) teachers' perceptions of the usefulness of alternative assessment strategies for grading student work (Grading); and teachers use of alternative assessment strategies in their classrooms (Assessment Use).

Data Analysis

The statistical techniques used for the teacher data include: (a) t-tests to examine pre to post mean differences within-Centers; and (b) Analysis of Covariance (ANCOVA) using the pretest scale score as the covariate to compare treatment groups (e.g., ADC vs. PDC teachers). The dependent variables were the post-SITACIP scale scores for assessing learning, assessment strategy use and grading. In the hierarchical framework, the Center is the independent variable. The Center is also a random factor because the sites were selected unsystematically from among all the potential sites to enable generalization of the results of the evaluation.

Results

The overall evaluation of SEPUP's impact on teacher change examined the differences between ADC and PDC teachers as well as between SEPUP and non-SEPUP (i.e., comparison) teachers (Roberts, 1996). The comparison teachers did not change on any of the measures used. The next section describes the differences between the ADC and PDC teachers on the selected SITACIP scales.

Quantitative Evidence

Table 1 summarizes the results on measures of teacher change related to their perceptions of the usefulness of alternative assessment strategies for assessing learning and grading as well as their actual alternative assessment strategy use and practice.

Insert Table 1 about here

In the beginning, both ADC and PDC teachers perceived alternative assessment strategies to be useful. Once faced with the reality of using alternative assessments to assess learning, guide subsequent instruction and grade students, ADC teachers turned to the practical issues and concerns that they faced in their classrooms daily, such as scoring large numbers of student papers and converting scores to grades. Local assessment moderation was the main support mechanism for teachers as they worked collaboratively to solve these practical dilemmas.

ADC and PDC teachers were significantly different on measures of assessment strategy use and the perceived usefulness of alternative assessment strategies for grading. These two groups were different on the Assessing Learning scale, but the difference was not statistically significant. However, the effect size is large, indicating that a large proportion of this difference is attributable to the Center. ADC teachers increased their use of open-ended questions and significantly decreased their use of closed-ended questions. Meanwhile, PDC teachers increased their use of closed-ended questions from the beginning to the end of the school year (i.e., they used traditional multiple choice tests) and reduced their use of alternative assessment strategies.

The results indicate that the PDC teachers did not grapple with the same practical and conceptual issues about assessment as their ADC counterparts. The results clearly underscore the rhetoric versus the reality in the SEPUP teachers' perceptions of alternative assessment and their classroom practice. The rhetoric of assessment reform indicates that alternative assessment is good, so the PDC teachers provided "socially desirable" responses (Nachmias & Nachmias, 1981, p. 222), and yet continued to use traditional assessment approaches. However, the reality of assessment reform is that it is fraught with problems and difficulties that may slow a teacher down in covering the required curriculum, anger parents who want to see their children taught in the same manner that they were, frustrate students who are used to an A through F grading system, and so on. The ADC teachers came face to face with these very real dilemmas of teaching with innovative materials and alternative assessments, so their perceptions of the usefulness of

alternative assessments were altered by the reality of implementation. ADC teachers re-assessed the usefulness of alternative assessment strategies through practice and reflection and collegial interaction -- which resulted in a small, non-significant decrease in their perceptions of the usefulness of such strategies for assessing learning and a statistically significant decrease on the Grading scale. Regardless of difficulties encountered, the ADC teachers remained committed to alternative assessment strategies (Roberts, 1996).

Qualitative Evidence

The level of success with implementation of local assessment moderation was determined to a large extent by the organizational context; in other words, the Center mattered. The features of the Assessment Development Centers (ADCs) that mattered were identified and organized using a case-ordered matrix approach (Miles & Huberman, 1994). The features are: quality and strength of leadership; institutional support for teacher professional development; and teacher proximity and collaboration. Figure 1 summarizes the qualitative evidence for the four ADCs, which are ordered from most to least successful with local assessment moderation. Success was measured in terms of: (1) completion of field test requirements by teachers within each ADC, such as number of activities moderated and selection of student exemplar papers; (2) teachers' reported levels of growth based on personal interviews and interviews with their principals; and (3) researcher observations of the moderation process corroborated by member checks with the ADC directors. Two of the six ADCs were dropped from the cross-level exploratory analysis because they did not fulfill their obligations as field test sites.

INSERT FIGURE 1 ABOUT HERE

Example of a Cross-Level Exploratory Analysis

The cross-level exploratory approach outlined below provides a view of teacher change across ADCs, thus contributing to an understanding of within-Center differences. This approach is best used when sufficient qualitative information is available in addition to quantitative

information, therefore a mixed-method approach was the most appropriate evaluation design in this context.

In a multisite situation such as the evaluation of SEPUP, there exists the potential for a program-by-site interaction (Sinacore & Turpin, 1991), so that simply pooling data to determine an overall program effect may lead to flawed interpretations (Keppel, 1991). The preliminary step in the cross-level exploratory analysis is to inspect for such interactions. To do so, the covariate means and the dependent variable (post score) means were plotted for each ADC. In order to illustrate the cross-level exploratory analysis, we present the results for the Assessment Use scale (for other scale results see Roberts, 1996). The Assessment Use scale was chosen because there was a significant interaction; three of the ADCs had positive slopes, one much steeper than the other two, and the fourth had a fairly steep, but negative slope.

Figure 2 presents the mean difference change plot for the Assessment Use scale. The mean difference in logits for each ADC is noted in parentheses along the horizontal axis. The ADCs are arranged from least (ADC 1) to most (ADC 4) successful with local assessment moderation based on the qualitative evidence presented earlier. Lines are used to connect the mean differences to indicate the pattern of differences across ADCs. The overall mean difference for the four ADCs is represented by the dotted horizontal line and can be easily contrasted with the individual Center's mean differences.

INSERT FIGURE 2 ABOUT HERE

Only ADC 4 had a decrease on the Assessment Use scale. One teacher in ADC 4 had a very high pre-score which dropped more than one logit to just above the overall post mean. The other three teachers in ADC 4 began with scores closer to the pre-scale mean for all ADCs. Two of these teachers had only slight decreases, but the third decreased by nearly one logit.

The other ADCs all had increases on the Assessment Use scale ranging from .19 logit to .83 logit. Two teachers, fairly new to teaching, in ADC 3 had very large gains on the Assessment Use scale, indicating that they reported using alternative assessment strategies much more

frequently than most other ADC teachers by the end of the field test. The overall post mean indicates that teachers most likely use alternative assessments about once a week, which sounds reasonable given the nature of the SEPUP assessment activities (e.g., lab reports).

Figure 3 presents the plot of mean scores for the four ADCs on the Assessment Use scale. The standard errors are indicated by the "whiskers." At the end of the year, the frequency of use of alternative assessments in ADCs 1, 2 and 4 is statistically about the same. However, ADC 3 is an outlier, which is explained qualitatively by the frequent assessment use reported by the two beginning teachers as noted above.

INSERT FIGURE 3 ABOUT HERE

Discussion

We believe this cross-level exploratory approach provides one solution to the interpretation of multilevel, multisite evaluation data when a purposive sample is too small to use sophisticated quantitative techniques, such as HLM. The example presented here indicates how the combination of quantitative and qualitative evidence can provide substantively interpretable information. Clearly, aggregating the quantitative evidence across all four ADCs would be misleading and discrepant cases would be masked. The qualitative evidence provides a basis for interpreting quantitative differences across as well as within-Centers in a meaningful way.

There were few significant changes at the teacher level, but teacher change takes time (Fullan, 1991; Little, 1993; Loucks-Horsley, Hewson, Love, & Stiles, 1998). Nonetheless, the results reflect the rhetoric versus reality of assessment reform in middle school science teachers' perceptions and practice (Roberts, 1996). On the one hand, those teachers engaged in using the assessment system reassessed the usefulness of alternative assessment strategies -- through practice, reflection and collegial interaction -- which resulted in decreases (rather than what one might have anticipated to be gains) on assessment-related teacher change variables. On the other hand, the treatment group of PDC teachers who used only the curriculum and did not engage in local assessment moderation retained rosy perceptions of the usefulness of alternative assessments,

but continued to use traditional forms of assessment. These results confirm what research has already suggested about teacher change: sufficient time must be allotted; implementation within one's own classroom must occur; inquiry and reflection must be promoted; and collegial discourse must be encouraged and supported (Barnett & Friedman, 1997; Guskey, 1986; Little, 1993; Loucks-Horsley, Hewson, Love, & Stiles, 1998; Richardson, 1990).

The organizational context factors that affected the level of success with the SEPUP Assessment System were identified through the qualitative analysis (Roberts, 1996). The Assessment Development Centers with strong local leadership, institutional support from the district and/or school level, and teachers who collaborated regularly were the most successful in implementing the SEPUP Assessment System. We concluded that district level support was the most conducive to success. Districts can support teacher professional development in various ways, such as financial resources or release time. Districts are more likely to implement reform activities than an individual school (Levine, 1995) and assessment reform once implemented is more likely to be sustained with district-level commitment (Chrispeel, 1997).

Reference List

Barnett, C., & Friedman, S. (1997). Mathematics case discussions: Nothing is sacred. In E. Fennema & B.S. Nelson (Eds.), Mathematics teachers in transition. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Burstein, (1980). The analysis of multilevel data in educational research and evaluation. In Review of Research in Education, 8, p. 158-233.

Caracelli, V.J., & Greene, J.C. (1997). Crafting mixed-method evaluation designs. In J.C. Greene & V.J. Caracelli (Eds.), Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass.

Chen, H-T. (1997). Applying mixed methods under the framework of theory-driven evaluations. In J.C. Greene & V.J. Caracelli (Eds.), Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass.

Chrispeel, J.H. (1997). Educational policy implementation in a shifting political climate: The California experience. American Educational Research Journal, 34 (3), pp. 453-481.

Fullan, M. (1991). The new meaning of educational change (2nd edition). New York: Teachers College Press.

Guskey, T.R. (1986). Staff development and the process of teacher change. Educational Researcher, 15 (5), p. 5-12.

Harmon, M. (1995). The changing role of assessment in evaluating science education reform. In R.G. O'Sullivan (Ed.) Emerging roles of evaluation in science education reform (pp. 31-52). New Directions for Program Evaluation, Number 65. San Francisco, CA: Jossey-Bass Publishers.

Ingvarson, L. (1990). Enhancing professional skill and accountability in the assessment of student learning. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. 327 558)

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall Inc.

Levine, H. (1995). The district as the fundamental unit of change. In St. John, M., Levine, H., Håkansson, S., Lopez-Freeman, M., Roberts, L., Shattuck, J., & Von Blum, R., A Study of the California Dwight D. Eisenhower Mathematics and Science Education State Grant Program: Lessons Learned about Designing Professional Development Projects, (Essay III). Inverness, CA Inverness Research Associates.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher, 23 (9), 4-14.

Little, J.W. (1993). Teacher professional development in a climate of educational reform. Educational Evaluation and Policy Analyses, 15 (2), pp. 129-151.

Loucks-Horsley, S., Hewson, P.W., Love, N., & Stiles, K.E. (1998). Designing professional development for teachers of science and mathematics. Thousand Oaks, CA: Corwin Press.

Nachmias, D., Nachmias, C. (1981). Research methods in the social sciences. New York, NY: St. Martin's Press, Inc.

National Research Council. (1996). National science education standards. Washington, D.C.: National Academy Press.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. Educational Researcher, 19 (7), pp. 10-18.

Roberts, L.L.C. (1996). Methods of evaluation for a complex treatment and its effects on teacher professional development: A case study of the Science Education for Public Understanding Program. Unpublished dissertation, University of California at Berkeley.

Roberts, L., Sloane, K., & Wilson, M. (1996). Local assessment moderation in SEPUP. Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Roberts, L., Wilson, M., & Draney, K. (1997). The SEPUP Assessment System: An overview. University of California, Berkeley: BEAR Report Series, SA-97-1.

SEPUP. (1996). SEPUP: Science Education for Public Understanding Program (pamphlet). Berkeley, CA: SEPUP, Lawrence Hall of Science.

Sinacore, J.M., & Turpin, R.S. (1991). Multiple sites in evaluation research: A survey of organizational and methodological issues. In New Directions for Program Evaluation, 50, pp. 5-18. San Francisco: Jossey-Bass.

Tobin, K. (1995). Teacher change and the assessment of teacher performance. In B.J. Fraser & H.J. Walberg (Eds.), Improving Science Education. Chicago, IL: University of Chicago Press.

Wilson, M. (1994). Community of judgment: A teacher-centered approach to educational accountability. In Office of Technology Assessment (Ed.), Issues in Educational Accountability. Washington, D.C.: Office of Technology Assessment, United States Congress.

Wilson, M., & Adams, R.J. (1996). Evaluating progress with alternative assessments: A model for Chapter 1. In M.B. Kane (Ed.), Implementing performance assessment: Promise, Problems and Challenges. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Table 1

Summary of t-tests and ANCOVA Results for Selected SITACIP Scales

Scale	t-test						ANCOVA		
	ADC			PDC			ADC vs. PDC		
	n	Post[a]	M Diff.[b]	n	Post	M Diff.	df[c]	F[d]	Adj. R^2
Grading	10	0.52	-.96*	12	1.46	.82*	19	9.39**	.49
Assessment Use	19	0.68	.20	11	.04	-.34	27	8.30**	.31
Assessing Learning	10	1.00	-.22	12	1.40	.47*	19	3.34+	.18

[a] Post: Post-test scale score in logits.

[b] M Diff.: Mean difference pre to post in logits.

[c] Degrees of Freedom associated with the S within group error.

[d] F ratio.

+ $p = .08$.

** $p < .01$.

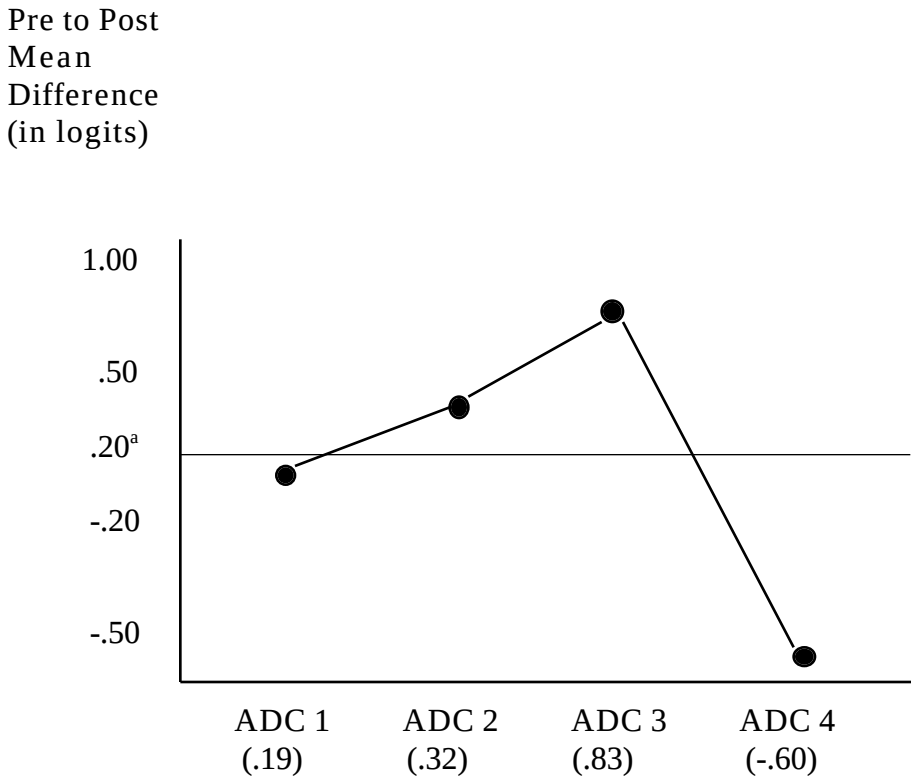
Figure Captions

Figure 1. Level of success with local assessment moderation by features of the Assessment Development Centers

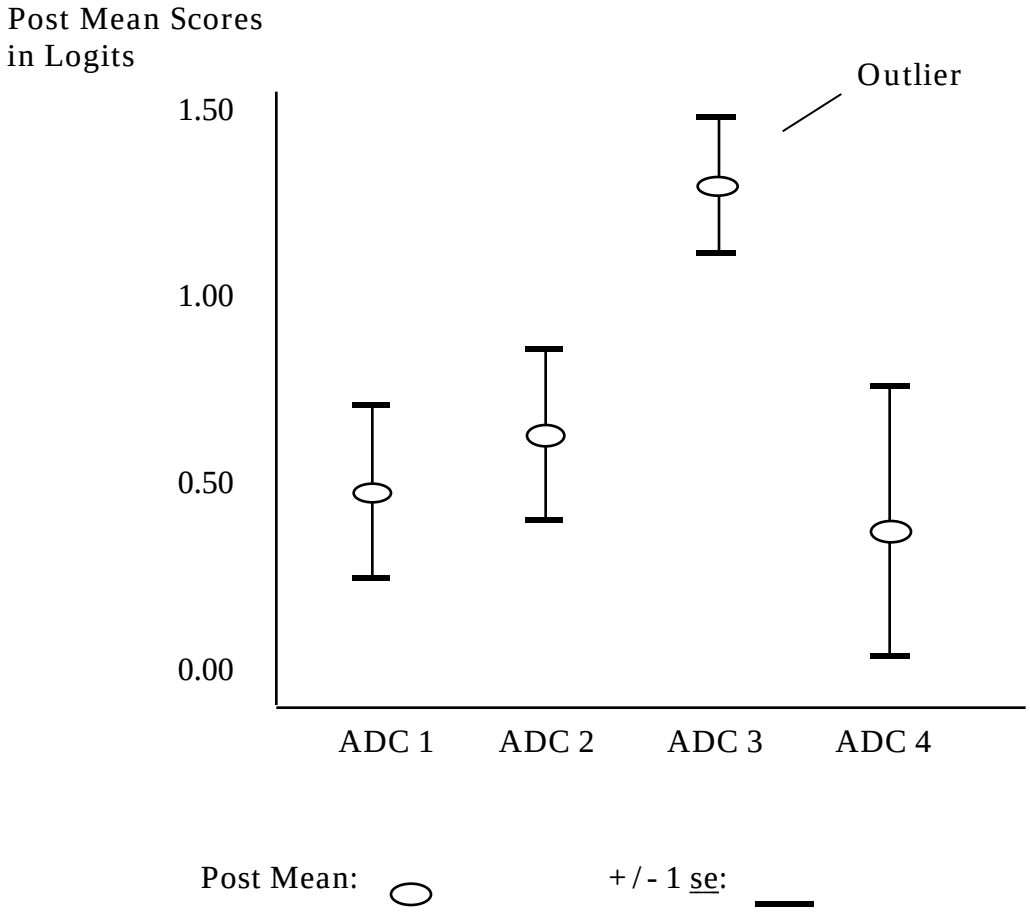
Figure 2. Change plot for the mean differences on the Assessment Strategy Use scale for the ADCs

Figure 3. Standard error plot for the post means on the Assessment Strategy Use scale by ADCs

Level of Success with Implementation of Local Assessment Moderation	Leadership	Institutional Support	Teacher Proximity and Collaboration
Very High (ADC 4)	Strong leadership of ADC director. Former SEPUP teacher functioned as moderation facilitator.	Single district with district administrator as ADC director. Principals supported teacher release time. Efforts parallel State mandates for change.	All four teachers in the same district, but at different schools. Teachers work on other projects together. Teachers communicate via electronic mail or telephone.
High (ADC 3)	Teacher leader in place of ADC director who took on other duties.	Two schools in a single district. Principals supportive of staff development efforts.	Teachers paired at the two schools. Beginning teachers were supported by experienced teachers.
Moderate (ADC 2)	Two districts involved. Two teachers co-facilitated moderation meetings. The ADC director missed a few meetings due to other district obligations.	Districts emphasizing assessment as staff development focus. Have several federally-funded science education reform projects in the State.	Teachers who co-facilitated were accepted in this shared role by the other teachers. Teachers called others in the same district between meetings (otherwise long distance).
Limited (ADC 1)	Five teachers from five different districts. Limited by the fact that the ADC director was also a classroom teacher.	Varying district mandates diluted level of success. Principalsí support varied, and there was no apparent district support. State drafting standards for science (K-12).	Limited somewhat by physical distance. Limited collaboration; called someone if a problem arose. Voice mail established by ADC director to provide weekly updates, but was not used by all teachers.



[a] Overall mean difference for all four ADCs. Mean Difference: ●


Post Mean Scores
in Logits
Outlier
1.50
1.00
0.50
0.00
ADC 1
ADC 2
ADC 3
ADC 4
Post Mean:
+/- 1 se: