DOCUMENT RESUME

ED 402 322 TM 025 819

AUTHOR Roberts, Lily; And Others
TITLE Local Assessment Moderation in SEPUP.
SPONS AGENCY American Educational Research Association, Washington, D.C.; National Science Foundation, Arlington, VA.
PUB DATE Apr 96
CONTRACT MDR9252906; NSF-RED-9255347
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Associ3tion (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Collegiality; Cooperation; *Educational Assessment; Field Tests; Intermediate Grades; Interprofessional Relationship; Junior High Schools; Middle Schools; *Professional Development; Program Implementation; Science Education; *Scoring; Student Evaluation; *Test Interpretation; Test Results
IDENTIFIERS *Consensus Moderation

ABSTRACT

Assessment moderation is a procedure in which scorers or raters meet to achieve a consensus on scores assigned to student work. In the Science Education for Public Understanding Program (SEPUP), local teams of teachers met regularly at six sites nationwide to score student work, review methods of assigning scores, discuss and resolve discrepancies in scoring, and reach consensus on exemplars of work for each score level. The sites were called Assessment Development Centers and were located in Alaska, California, Colorado, Kentucky, Louisiana, and Oklahoma. Moderation sessions thus served purposes related to technical aspects of assessment, although SEPUP moderation sessions served additional purposes and provide additional benefits for teachers. Because moderation was part of a field test, teachers in the local group shared common concerns and experiences. As they field-tested a new approach to middle school science instruction, they were learning a new approach to assessing student performance. The moderation process went beyond its traditional purposes to purposes of professional development and teacher collegiality. This discussion of SEPUP local assessment moderation explores the way moderation meetings were intended to function and the ways they did function in reality. Preliminary insights into factors influencing successful implementation are proposed, and issues in the role of local moderation in assessment and professional development are discussed. An appendix summarizes the moderation process. (Contains 2 figures and 42 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
LILY ROBERTS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 402 322

Local Assessment Moderation in SEPUP

Lily Roberts
University of California, Berkeley

Kathryn Sloane
University of Illinois at Urbana-Champaign

Mark Wilson
University of California, Berkeley

March 1996

Paper presented at the annual meeting of the American Educational Research Association, April 9, 1996, New York, NY.

TM025819

BEST COPY AVAILABLE

ERIC

Acknowledgments

This paper is the result of a united effort of many individuals who have worked with us directly or cooperated with our efforts tangentially. We would first like to thank all the Science Education for Public Understanding Program (SEPUP) teachers, their students, and principals who were on the "frontline" in our efforts to develop and field test the embedded assessment system that is now an integral part of the *Issues, Evidence and You* course. Particularly, we wish to acknowledge those individuals in SEPUP Assessment Development Centers (ADCs), without whose commitment to the task of local assessment moderation, this paper would not have been possible. The ADCs were the window through which we were able to see how the assessment system unfolded in the classroom. Without the valiant efforts of the ADC teachers and the center directors, we would be lacking much of the richness of spirit and detail about local assessment moderation that we have tried to capture in this paper.

We also would like to acknowledge our SEPUP development colleagues at the Lawrence Hall of Science, led by Dr. Herbert D. Thier, for their support and cooperation as we worked to build an embedded assessment system that would fit seamlessly into the fabric of the *Issues, Evidence and You* course. Finally, we wish to thank our SEPUP Assessment Team colleagues that contributed to this paper in one way or another; they are Robin R. Henke, Megan Martin and Chris White.

This research was supported in part by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation and the National Center for Education Statistics (U.S. Department of Education) under NSF Grant #RED-9255347 Further, this project has been supported by NSF grant No. MDR9252906, and also by other support through SEPUP at the Lawrence Hall of Science. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

Abstract

Assessment Moderation is typically defined as a procedure where scorers or raters meet to achieve consensus on scores assigned to student work. In the Science Education for Public Understanding Program (SEPUP), local teams of teachers met regularly to score student work, review methods of assigning scores, discuss and resolve discrepancies in scoring, and reach consensus on exemplars of work for each score level. Moderation sessions thus served purposes relating to the technical aspects of assessment, specifically reliability and rater consistency in scoring. Traditionally, this has been the primary purpose of moderation.

In SEPUP, however, moderation sessions served additional purposes and yielded additional benefits for participating teachers. Because moderation was part of a field-test effort, the teachers in the local moderation group shared common experiences and concerns: they were field-testing a new approach to middle school science instruction, and they were learning a new approach to assessing student performance. Regular meetings with their colleagues provided a context for ongoing collegial support and a mechanism for constructing and reviewing their understanding of the course and the assessment system. Therefore, the moderation process went beyond the traditional purposes of technical sufficiency, toward purposes of professional development and teacher collegiality.

In this paper, we discuss the purposes of the SEPUP local assessment moderation process and explore the roots of these purposes in related research literature: moderation as an assessment procedure; teacher roles in assessment; and models of professional development. We then describe SEPUP local assessment moderation groups--the ways they were *intended* to function, and the ways they *did* function in various sites across the country. Preliminary insights into factors influencing successful implementation are proposed. Finally, we discuss issues that arose and future directions in the study of the role of local moderation in assessment and professional development.

In the assessment component of the SEPUP project, our aim was to build upon advances in assessment methodology to design a comprehensive, integrated assessment system for the middle school science course, *Issues, Evidence, and You*. The system is based on four principles. The first principle is that the system must take a *developmental perspective*. That is, the information flowing through the system must focus principally on the student developing through the year-long curriculum. The second principle is *instructional fidelity*. The assessments are to be an integral part of the teaching and learning process--to reflect the hands-on instructional approach and to reinforce the curricular goals of evidenced-based decision making. The third principle is that the system is under the *management and responsibility of teachers*. Teachers must be able to manage the system (including scoring and interpretation) efficiently and to use the student assessment results effectively within the context of classroom teaching and learning (see Sloane, Wilson & Samson, 1996, for a discussion of the principles and components of the full assessment system). The fourth principle is *high technical quality*. Assessments need to be valid and reliable indicators of student progress and performance on a defined set of constructs central to the course objectives[1]. Responses from different types of prompts (written responses, performance measures, oral presentations, etc.) must be integrated into an accurate and coherent assessment of student performance at any one point in time, and of student progress over the course of the year. Scoring procedures need to capture the complexity of the constructs, yet be efficient and reliable.

A host of assessment development and psychometric analyses issues had to be addressed in the process of designing the "nuts and bolts" of the system. But a predominant concern from the beginning of the project was translating the materials and procedures we developed "in the lab" to the "real world" of teachers, students, classrooms, and schools. The crux of the issue was two-fold: a) how to prepare teachers to understand and use alternative forms of assessment and state-of-the-art psychometric tools, and how to support them through this learning process; and b) how to work with teachers during the development phase to pilot-test emerging versions of the materials, in order to obtain the student data necessary for producing and modifying components of the system, and for evaluating (formatively) the curricular materials and assessment procedures.

The central mechanism we employed to meet these teacher enhancement and assessment development needs was *local assessment moderation*. The model of local assessment moderation we developed includes the basic elements of consensus moderation (a "scorer-calibration" process currently used in some state- and district-level alternative assessment programs), the general structure of site-based feedback meetings, and the emerging principles of collegiality, collaboration, and reflection inherent in new approaches to professional development. The local assessment moderation model may offer new perspectives on the roles of teachers in assessment reform, new techniques for enhancing teachers' professional learning and practice in assessment, and new methods for strengthening the linkages between curriculum, instruction, and assessment.

[1] The variables that formed the framework for the course and the assessment system in *Issues, Evidence, and You* are: Designing and Conducting Investigations; Evidence and Tradeoffs; Understanding Concepts; Communicating Scientific Information; and Group Interaction.

Foundations for the Local Assessment Moderation Model

In this section, we review three lines of research relating to: a) the role of moderation in alternative assessment; b) the roles of teachers in alternative assessment, and the professional development needs of teachers assuming these roles; and c) recent research on models and methods of quality professional development. While these literatures are distinct, we have combined components of them into our model of local assessment moderation as an integrative factor in assessment, which is described in the subsequent section.

Consensus Moderation as a Method of Quality Assurance in Alternative Assessments

Shifting from standardized tests to performance-based assessments has created a need to "rethink the criteria by which the quality of educational assessments are judged" (Linn, Baker, & Dunbar, 1991). The traditional concepts of validity and reliability of measures are still critical, but both have been recast as criteria for judging assessment in light of recent reform efforts in state and national tests (Baker, O'Neil & Linn, 1993; Messick, 1994, 1989; Moss, 1994) as well as in other nations' educational reform efforts (e.g., New Zealand Qualifications Authority, 1992).

Consensus moderation, where groups of judges come together to agree on a common standard using specific pieces of student work, is one means to ensure comparability of assessment results (Abbott, 1991; Baker, O'Neil & Linn, 1993; Ingvarson, 1990; Linn, 1993b, 1994; Richards, 1992; Wilson, 1992a, 1994). Moderation comes in many forms, such as defining criteria for assessment, external moderation, exemplars of student work, or common assessment tasks and reference tests (Harlen, 1994; New Zealand Qualifications Authority, 1992).

Consensus moderation is fairly new to education in the United States. Primarily, "judgmental scoring" has been used in the assessment of writing (Linn, 1993b). The movement toward performance-based assessments in high-stakes testing in the United States has initiated a scholarly examination of alternative methods for ensuring quality of assessments and comparability of results (Baker, O'Neil, & Linn, 1993; Linn, 1993a, 1993b, 1994; Linn, Baker, & Dunbar, 1991; Wilson, 1992b). Moderation has emerged in much of this literature as one means to ensure quality control. Other means of establishing comparability exist, such as equating, calibration and statistical moderation, and have been described elsewhere (Baker, O'Neil & Linn, 1993; Linn, 1993b; New Zealand Qualifications Authority, 1992; Wilson, 1994). These quantitative approaches to comparability are typically used for traditional tests, but have more recently been extended to performance assessments (Wilson & Wang, 1995).

The consensus moderation process was used in the state of Victoria in Australia in response to curriculum reform efforts in the 1970's. The reform movement led by teachers' organizations and unions was toward inquiry-based education and away from fact-based learning. Up to that time, teachers' grades of student work were not generally used for graduation or college admission purposes. Student selection into higher education was based primarily on a statewide subject-based examination taken at the end of year 12; the universities controlled the process. The consensus moderation process was implemented as a means "to ensure that teachers' standards and marks (grades) were comparable from school to school and teacher to teacher" (Ingvarson, 1990). Comparability was seen as a prerequisite for support of this innovation by parental, business and university groups.

Classroom teachers have been drawn into the consensus moderation process. In the U.S., consensus moderation groups convened for quality assurance in statewide assessments often include classroom teachers as scorers. For example, this was a common strategy used in scoring the California Learning Assessment System (CLAS) Science (California Department of

Education, 1995). In the U.K., as part of broad-sweeping reforms in certification and in assessments of the National Curriculum through the 1988 Education Reform Act, classroom teachers became more integrally involved in the process of scoring student work and establishing standards of student performance, as participants in the "reconciliation model" of moderation (Torrance, 1995a). The use of teams of teachers in these types of scoring processes has been a necessary logistical procedure to handle the increased demands of scoring that alternative or performance assessments produce. It has also had the effect of introducing a core group of teachers to these new assessments and related scoring procedures. There may also be an interest in stimulating teachers to try new assessment procedures in their classrooms, or to help teachers better understand the implications of alternative assessments for their curriculum and instruction (e.g., an emphasis on problem-solving). The success of this latter intent may be limited, however, because teachers are not scoring their own students' work, and because the "high stakes" tests are not linked directly to a specific course or curriculum.

Teachers' participation in consensus moderation has occurred primarily through these types of high-stakes assessment programs that focus on accountability and certification purposes. To our knowledge, there have been fewer instances of teachers participating in moderation of classroom-based assessments that focus more on instructional management and monitoring purposes.

Role of Teachers in the Alternative Assessment Process

There appears to be three primary roles recognized for teachers in the alternative assessment movement. The first, mentioned above, is participating in consensus moderation groups convened for high stakes assessment programs. This type of involvement is only tangentially related to the teachers' work in their own classrooms, however. And, a relatively small percentage of the total teaching force is directly involved in this activity. While participating teachers may discuss their experiences with their colleagues, the procedures used to develop and score alternative assessments may remain a mystery to those teachers who do not participate in these scoring groups.

The second role is in gathering materials for use in alternative assessment programs. For example, teachers may be asked to assist students in selecting materials for portfolio assessments, or they may select the materials themselves. As another example, the assessment program for the National Curriculum in the U.K. included a combination of "teacher assessments" of coursework and teacher-administered "standard assessment tasks" (Torrance, 1995b).

The third role is to ensure that students are experiencing the type of curriculum and instruction--in the classroom--that will be assessed. While not often explicitly stated, this role is certainly implicit in many of the arguments for alternative assessment programs. As Resnick and Resnick (1992) argued: "Assessments must be designed so that when teachers do the natural thing--that is, prepare their students to perform well--they will exercise the kinds of abilities and develop the kinds of skill and knowledge that are the real goals of educational reform." Torrance (1995a) summarized the rationale for assessment-driven educational reform as follows: "broaden the scope of the assessment system and increase the complexity and demands of tasks involved, and you will broaden the curriculum and raise the standards of teaching. . . .In turn, these new authentic assessments will lead to improved teaching and learning in schools as teachers adapt their curriculum and laboratory procedures to ensure that their students succeed at the new tasks" (p. 3). Certainly, teachers are increasingly being held accountable for their students' performance on high-stakes alternative assessments and no doubt feel pressured to adapt their instruction--and perhaps their classroom testing procedures as well-- to prepare their students for these assessment demands. In some states, there are formal staff development programs to assist teachers in adapting their classroom testing procedures to better match those that students will encounter in statewide tests. For example, as part of the Kentucky Education Reform Act (KERA), workshops were offered in constructing open-ended items, on methods of selecting student work for portfolios, and the like.

Using alternative assessment practices in the classroom requires teachers to develop new classroom management strategies, a new knowledge base about assessment practice, and new attitudes about the purposes of student assessment (Chittenden, 1991; McCallum, Gipps, McAlister, & Brown, 1995; Torrance, 1995b; Zessoules & Gardner, 1991). Little (1993) has noted that "at the local level, teachers' interest in alternative forms of assessment far exceeds their professed skill in constructing, evaluating, or incorporating them into their practice. Further, teachers do not have adequate resources available to them from the research and test development communities." While it is recognized that teachers need opportunities to learn and practice new assessment techniques if they are to incorporate these into their own classroom instruction (Harmon, 1995), systematic and concerted efforts to determine effective methods for professional development in assessment are still rare. And researchers are recognizing that the process takes time.

In one classroom-based assessment project, Shepard and her colleagues (Shepard, 1995) worked with a small group of third grade teachers to use performance measures in reading and mathematics in their classrooms. After a year of intensive work in that project, the teachers became more "sophisticated" about scoring criteria (e.g., focusing on the intended construct, and recognizing the possible multiple dimensions in scoring a performance task). The researchers concluded, however, that "current calls for assessment driven reform acknowledge the need for staff development but tend to underestimate the extent and depth of what is needed" (p. 42). They proposed that "what is needed" includes: (a) appropriate materials to try out and adapt; (b) time to reflect and develop new instructional approaches; and (c) ongoing support from experts to learn (and challenge) the conceptual bases behind intended reforms.

Emerging Models of Professional Development

Shepard's (1995) conclusions about the time and ongoing support teachers need to learn new assessment strategies are consistent with current insights regarding successful professional development experiences in general (Ross & Regan, 1993). In the professional development literature, researchers and educators acknowledge the severe limitations of the "one-shot approach" (Fullan, 1982). It takes *time* for teachers to construct an understanding of new techniques and to practice ways to incorporate these into their existing instructional strategies--or to change their instructional approach more dramatically. Attention to "length of time" is based on the recognition of two factors: (a) constructing understanding is a *process*, and (b) teachers must engage in this process *while* they are continuing their normal (and demanding) work in classrooms (Loucks-Horsley & Stiegelbauer, 1991). Therefore, there is an increasing emphasis on the need for *ongoing* opportunities for teachers to learn and reflect, while they are engaged in the process of practicing new techniques in their classrooms (Guskey, 1986; Loucks-Horsley et al., 1989; Shepard, 1995).

Additional principles of quality professional development practices have also been proposed. Little (1993), for example, proposes the following principles as standards for judging the quality of a professional development program: (a) "offers meaningful intellectual, social, and emotional engagement with ideas, with materials, and with colleagues both in and out of teaching;" (b) "takes explicit account of the contexts of teaching and the experiences of teachers;" (c) "offers support for informed dissent;" (d) "places classroom practice in the larger contexts of school practice and the educational careers of children;" (e) "prepares teachers (as well as students and their parents) to employ the techniques and perspectives of inquiry;" and

(f) "balance support for institutional initiatives with support for those initiated by teachers individually and collectively" (p.138-139).

More attention is also being given to the importance of collegiality, collaboration, and community in teachers' professionalism and professional growth. For example, in examining the role of Urban Math Collaboratives in teachers' professional knowledge, competence, and commitment, Little and McLaughlin (1991) describe the principle of colleagueship: "the idea that the professionalization of teaching and the improvement of teaching will be advanced more surely when colleagues join together on matters of professional practice" (p. 18). They list three themes emerging from this principle in these Collaboratives. First is the opportunity for collegial exchange. This opportunity includes first, blocks of time (outside of the school day) for collegial exchange. But to make the most of this time together, teachers must have a shared, common purpose to their time together, resources of knowledge and expertise to ensure their joint work is good, and shared norms and values that favor open discussion and debate about matters of professional practice. It is "opportunities of this sort [that] distinguish the Collaborative from conventional professional development" (p.19). A second theme is that mutual support is accompanied by mutual obligation. And the third theme is that teachers construct rigorous standards for colleagueship (i.e., what makes a good colleague).

As described above, there are models of consensus moderation that appear to be working satisfactorily as a mechanism for establishing quality control in alternative assessment programs. State and district wide assessment procedures, however, do not address the critical issue of how to translate assessment reform into classroom practice. Clearly, issues relating to teachers' professional knowledge and practice in assessment will have to be addressed more directly if teachers are to effect meaningful change in assessment in their own classrooms. New models of professional development (construed broadly) may offer some effective principles for designing professional development activities in assessment.

Local Assessment Moderation as an Integrative Factor in Assessment

We have developed a model for combining the benefits of consensus moderation and of effective professional development principles which we call *local assessment moderation.* In local assessment moderation, teachers within a defined geographic area who are teaching a specific curriculum meet regularly to score, interpret, and discuss their students' performances on common assessment tasks. There are two distinct, yet interrelated purposes from the teacher's perspective, and one from the developer's perspective.

The first purpose is enhancing the *technical quality* of the teachers' procedures for assessing student work. Teachers score a combined, common set of student work and gain practice using and understanding the scoring procedures. By discussing commonalities and discrepancies of their scores, teachers gain a deeper understanding of the scoring levels and standards of performance for each level. By reaching consensus on the scores a given piece of work should receive, the teachers "standardize" their interpretation of the scoring guides and become more consistent in their own scoring over time, as well as more consistent as a group in their ratings at any one point in time. Further, the teachers collaborate on adapting the scoring procedures or standards of performance to fit local needs and conditions (for example, by relating the course scoring guides to the rubrics used in district- or state-wide assessment programs). The aim is a set of teacher-scored measures that meet technical standards of reliability (consistency) and validity (what the scores mean in terms of student understanding and performance). These measures can then be used in powerful--and defensible--ways in subsequent interpretations (such as maps of performance over time) or evaluation studies (conducted, for example, by the teacher herself for course improvement or accountability, or by administrators, curriculum developers, or researchers for more summative evaluation purposes).

The second major purpose of local assessment moderation is to *support teachers' professional learning and practice* in student assessment and in their general approach to teaching. Making substantial (and substantive) changes in assessment practices requires learning new concepts and techniques, trying these out in classroom practice, reflecting on the results, and generating new plans for managing and integrating these techniques in instruction. Local assessment moderation groups provide a structure for learning, experimenting, and reflecting; they provide time for teachers to share ideas, analyze problems, and plan strategies; and, they provide a support network of colleagues with similar interests, aims, and experiences. As teachers focus on the work at hand, and share ideas and insights into the process, the opportunity emerges to grapple with "larger" or more conceptual issues in student assessment, such as: what it means to assess student progress and to assign and interpret scores; how to use assessment information in the process of assigning grades; how to explain the assessment process to students, parents, administrators, and other audiences; how to give more meaningful and useful feedback to these audiences about how students are progressing and what their scores and grades really mean; and, perhaps most importantly, how assessment can be used to plan and enhance instruction.

Finally, in a curriculum and assessment development context, local assessment moderation groups can achieve a third purpose: *feedback to developers*. In the moderation sessions, teachers discuss the assessment task, the types of student responses generated by the task, and the "fit" of the responses to the scoring guide. These topics can emerge "naturally" from the scoring activities and consensus-building discussions that takes place in moderation. Teachers' (and students') reactions to the assessment materials (task, scoring guide, etc.) provide invaluable feedback to developers on how well the materials are working and on the placement of the assessment prompts in the flow of instructional activity. "Exemplars" of student work identified by the teachers for each score level provide feedback to developers on the type of student learning that is occurring. Further, exemplars provide "real" examples of student response that can be used to interpret student scores and serve as "training examples" for other teachers learning the system.

These "purposes" of local assessment moderation represent an ideal. In SEPUP, we experimented with the process of local assessment moderation during the two years of national field-testing of the curriculum and assessment materials. In the following sections, we describe the "working model" we designed and the ways in which local sites implemented the model in practice. From information obtained through site visits, teacher interviews, and staff observations of the moderation sessions, we report preliminary findings on the results of local assessment moderation relative to the purposes listed above--the areas of both promise and problems. Finally, we offer some reflections on what we have learned and what we believe are the future steps in fully understanding the potentials and limitations of this model of local assessment moderation.

Local Assessment Moderation in SEPUP

Classroom teachers who are implementing the SEPUP assessment system meet in groups to discuss the assessment tasks, to jointly score student papers, to reach consensus on scoring procedures and standards of performance, and to discuss the implications of student performance for instruction. In the SEPUP field trials, teachers at six sites nationwide met regularly to moderate 10 of the embedded assessment tasks. The sites were called Assessment Development Centers (ADCs), and were located in the following States: Alaska, California, Colorado, Kentucky, Louisiana, and Oklahoma. This section describes an idealized account of the local assessment moderation model, how teachers were prepared to use the moderation process and SEPUP moderation groups in practice. The latter is interwoven with comments from teachers who participated in moderation.

Description of the Local Assessment Moderation Process

The first meeting of a group of teachers, especially if they are not very familiar with each other, is best spent ensuring that the group becomes acquainted and that sufficient preparation is provided about the moderation process and the assessment system in general.

After the initial moderation meeting, a typical moderation (a synthesis of best practices from all ADCs) would be carried out as follows:

1. *Teachers' would have discussed the prompt and the scoring guide for the appropriate variable(s) at the previous meeting.* They would reach a "consensus of interpretation" before implementing the assessment task. Once the task was implemented in their own classes, they would score the papers using the Scoring Guide, and select a set of papers that exemplify a range of responses as well as difficult to score papers.

2. *Teachers would bring Xeroxed copies of a range of student papers on the same assessment task agreed upon at the prior meeting.* Sharing papers in advance or scoring round-robin (not Xeroxing, but simply passing along to the next teacher) were not effective for various reasons, including: geographic distances between some teachers; teachers lagging behind others and not completing the designated assessment until the day before the moderation meeting; and teachers preferred to take back examples from others' classes to share with their own students as exemplars of different score levels.

3. *"Comment without comments" was an adaptation by the Kentucky ADC of the Scoring Guide Discussion.* Teachers were able to "vent frustrations" with the scoring guide, the prompt, the activity, or whatever. Each teacher makes a brief comment that the moderation leader records. Teachers are not supposed to piggy back on each others' comments, but rather to get a range of concerns out on the table.

4. *Teachers would then distribute their 4-6 papers to the others in the group.* Students' names were replaced by codes to ensure anonymity, such as the teacher's or the school's initials and sequenced numbers (e.g., R1, R2, R3, S1, S2, S3...). Teachers would then read and score the other teachers' papers, having scored their own before they arrived in order to select a subset to bring to moderation. After scoring, the scores are compiled on a chalkboard, a transparency or newsprint, so that all can see.

5. *Moderation to reach consensus would then proceed.* The moderation leader moves the group toward consensus.

6. *After moderation.* The teachers select exemplar papers of the different score levels from among the papers they have scored. Teachers can complete a moderation reflection form to help them synthesize their thoughts about the instructional implications that this process has had for them. There needs to be some kind of general discussion at the end of the meeting to bring closure to the process and to allow the teachers to debrief on SEPUP, moderation, instruction, or management issues. Teachers may wish to return to a discussion of some of the issues identified during the "comment without comments" exercise, or they may wish to address specifics about the

scoring guides. The moderation leader needs to be sure that the group leaves the meeting with a "consensus of interpretation" about the next assessment activity to be moderated, including the variable(s) and element(s) to be assessed (see Appendix for a summary).

An Exemplary Moderation Interchange

The following is a fabricated example based not on an actual moderation, but rather a synthesis of what occurred in a number of the moderations over the year. ADCs audiotaped moderation sessions and sent these tapes to the Assessment Project. This information combined with other moderation feedback and observations of moderation sessions during site visits provided the basis for this constructed dialogue.

Figure 1 presents a hypothetical collection of pre-moderation scores for four teachers. The moderation leader would note the student papers for which there are different scores and then lead the group in discussion to reach consensus on these student papers.

Figure 1. Pre-moderation, initial scores by teachers

Students	Ms. S	Mr. R	Mr. W	Miss Z
S1	2	3	2	2
S2	3	3	3	2
S3	2	2	2	2
R1	1	1	1	1
R2	2	0	1	1
R3	1	2	1	2
...	---	---	---	---
Z3	2	2	2	3

Moderation Leader: "Mr. R why do you feel S1 deserved a 3? Everyone else scored this paper a 2."

Mr. R: "Well, I was wavering between a 2 and a 3, but the presentation was well done, so I thought the extra effort pushed it over the edge to a 3."

Ms. S: "Mr. R is correct that the paper is nicely put together, but we're not scoring the *Communication* variable, we're scoring on the element *Using Evidence*. So this student's paper is a 2 on *Using Evidence*. But I agree, I would give it a 3 on Communication."

Moderation Leader: Mr. R, do you agree with Ms. S on this paper, or did you have other points specific to the *Evidence and Tradeoffs* variable?

Mr. R agrees and the moderation leader moves on to the next student paper, S2, which is also easily resolved once Miss Z agrees that she had a different interpretation of this students' paper at first.

Moderation Leader: "Well, R2 is obviously one that needs some discussion. Who would like to start?"

Ms. S: "I better take a closer look at this paper, but I don't think it's a zero. I thought the student showed some use of evidence in the paper. Let me look at it again."

Mr. W: "This paper is clearly a 1. R2 only provides *subjective reasons* for his decision, even though he draws some remarks from the activity, his reasons are off-the-mark."

Mr. R: "What's the difference between *subjective* and *irrelevant*? Yes, these are subjective reasons, but as you say they're off topic, so they're irrelevant. I still think this paper is a ZERO."

After several more minutes of heated discussion followed by a quick re-read of the brief paper as directed by the moderation leader, the teachers reach agreement with Mr. R. The consensus score for student R2 is changed to zero. This type of scenario was presented by a Kentucky ADC teacher as one of the values of the moderation process. She felt that even if her score was far different from all the other teachers, she could still argue her case, and sometimes persuade the group to agree with her and other times not. She felt a strong sense of validation and power as a teacher from this type of experience.

The moderation leader then proceeds through the balance of the student scores in like manner to reach consensus. At the end, Miss Z does not feel that her student Z3 scored less than a 3, but she concedes to the group consensus of 2. However, she notes that she will retain the score of 3 for her own grading purposes.

Figure 2 presents the final product of the moderation of one element of one variable. In most cases, the moderation would proceed with the next element of a variable or perhaps move onto a second assessment activity.

Figure 2. Moderated, consensus scores

Students	Ms. S	Mr. R	Mr. W	Miss Z
S1	2	~~3~~ 2	2	2
S2	3	3	3	~~2~~ 3
S3	2	2	2	2
R1	1	1	1	1
R2	~~2~~ 0	0	~~1~~ 0	~~1~~ 0
R3	~~1~~ 2	2	~~1~~ 2	2
. . .	. . .	. . .	. . .	. . .
Z3	2	2	2	~~3~~ 2

Preparation to Use Moderation

In August 1994, a dozen SEPUP teachers were convened to revise scoring guides and to prepare assessment materials for the ADCs to use for the field test. These teachers were prepared to use the moderation process through a mock assessment activity using a scripted outline for a moderation meeting (see Appendix A). These individuals were then responsible for preparing the other teachers in their ADCs in the use of assessment moderation. This process was modified to meet the needs of teachers within the various centers, and these modifications were considered when revising the description of the assessment moderation process for the assessment manual. Each center had an Assessment Project liaison who attended the first moderation sessions of the year in an effort to provide additional support to ADC teachers as they learned how to "moderate."

Participants' Reactions to the Local Assessment Moderation Process

Participating in the local assessment moderation process helped SEPUP teachers achieve a conceptual understanding of the assessment system. It also provided a forum in which teachers could share difficulties and challenges with colleagues and obtain useful ideas, or have an opportunity to openly brainstorm solutions. Interestingly enough, we found that moderation fulfills the needs of both teachers who feel isolated in their own school contexts as well as those that already have ongoing collegial relationships at school. Moderation was described as a "leveling process" in one ADC, as a "normalizing experience" in another, and as a process for developing and fine-tuning standards in a third. In other words, when teachers came together as a group they were able to look at student performance through many eyes and yet come to a shared understanding. Furthermore, the process honored the input of beginning teachers as much as that of 25 year veterans.

Interviews with participating teachers and ADC directors, yielded some interesting anecdotal information about the value of moderation. The following comments[2] summarize the types of benefits teachers (and ADC directors) felt they received from participating in ongoing local assessment moderation at their sites.

Some comments related to the role the local assessment moderation process played in providing collegial support for understanding the assessment system.

> Building an understanding of embedded assessment and using it effectively. One teacher noted that although she receives strong collegial support at her school, she feels that moderation with the other ADC teachers was important to help her understand the scoring guides and how to apply them to specific lessons. She said that she feels that the moderation process helped her use the scoring guides effectively, with the practice making it easier to score each time[3].
>
> Benefiting from contributions of colleagues. Another teacher from the same ADC, said that the moderation process was very helpful to her because she feels so isolated in her school (she teaches all 7-8 grade math and science classes, so she is the department). She noted that she always felt good after moderation. She described moderation as "six minds rethinking" how things went and the awareness that was brought about through "exchange of thoughts." She believes the process is helpful because it "broadens thinking about assessing student learning." She gained ideas from the other SEPUP teachers about how to modify instruction or things to enhance instruction. She co-facilitated the moderation sessions with another teacher, and yet she places herself as an equal with the others in the group when it comes to learning and benefiting from the process.

The focus on scoring student work and reaching consensus ensures that teachers receive regular feedback on the student learning process (Guskey, 1986). The moderation process allows teachers to engage in the collaborative investigation of important problems of practice and to share expertise (Hollon, Roth, & Andersen, 1991). An important outcome of the moderation process is the attention to instructional implications raised by the assessment activities.

[2] The comments are paraphrased from field notes and audio tapes of interviews and focus groups.

[3] The researchers have complete records of who said what in these sessions. We are not providing sources in this paper in order to preserve anonymity.

Feedback on Student Learning. One teacher reported that she has been able to realize through moderation that she "hasn't gotten through" to her students about *designing investigations* and that she expects too much of the students sometimes. She indicated that the moderation process allows "time to reflect on what you did right or wrong" and that you have "the opportunity to listen to others and get ideas to use next time."

Another teacher noted that he took a different viewpoint about his students' capabilities after engaging in moderation. He feels that he has been given a more "global viewpoint" of what student learning is actually taking place. He reports that he is able to see general trends in growth rather than task specific, and he is able to determine if his students can apply what they have learned.

Another important outcome of moderation is the opportunity to reflect on one's own learning. Several teachers noted the importance of the moderation process as a time to reflect or review their own teaching.

Promoting reflective practice. One teacher noted that she feels that the moderation process has contributed to her professional growth because it makes her examine her teaching and students' learning. She adds, "you reflect on your purpose as a teacher: to see students grow and change, and have students see that they're learning something that's applicable." She reported that she sees that her students are referring back to prior knowledge and applying it to something else.

The assessment moderation process also serves as a medium for feedback to the teachers and the broader learning community. The teachers gain information on how their application of the scoring guides aligns with that of their peers. In addition, information is communicated back to the SEPUP Assessment Project about the assessment activities, the scoring guides, and the SEPUP assessment system as a whole. This information was used for the formative evaluation of the assessment system. Several teachers also engaged in professional activities, such as conference presentations about their SEPUP experience.

Comparing your students to others. This can be an ugly can of worms or it can meet the measurement objective of ensuring comparability and quality control. In one ADC, this issue led to strong defensive posturing on the part of some teachers, and unfortunately this group had a less than successful experience with moderation. In other ADCs, the opportunity to compare was embraced as an opportunity to clarify one's expectations of students or "to see how I can bring my students up to a certain level." In the ADC that we agree has had the most success with moderation for a variety of reasons, the teachers feel that the moderation process enabled them to become better at evaluating their students.

Communicating with a broader audience. The ADCs communicated with other audiences about moderation. In addition to sending compliance information back to the Assessment Project, the ADCs used other forums to share their SEPUP experiences. Several ADCs participated in presentations at their state Science Teachers Association conferences. Teachers made presentations to their own schools and/or districts about the SEPUP assessment system. Others made connections with their state education departments and those involved in developing or implementing state science standards.

Moderation does not do away with the "systemic" issue of grades still being valued by parents, school boards and universities. Consequently, one of the major psychological hurdles for SEPUP teachers using scoring rubrics has been the conversion of scores to grades.

Discussion of Preliminary Results

An ongoing evaluation of the implementation of the consensus moderation process in Victoria was conducted from 1981, when it was first implemented, to 1984, with a follow-up study in 1989. Based on the four year evaluation in Victoria, Ingvarson (1990, p. 9) noted "the importance of regarding moderation as a complex innovation requiring a considerable period of time for [teachers'] learning and unlearning during its implementation." The Victoria study indicates that the consensus moderation process had "impressive side effects on the professional development and accountability of teachers" (Ibid., 1990, p. 2).

Many of the comments from the SEPUP ADC teachers are consistent with the findings of the Victoria evaluation (Ingvarson, 1990). In the Victoria evaluation, it was found that involvement in the consensus moderation process: (a) added significantly to teachers' skills for assessing student learning; (b) enhanced teachers' ability to evaluate and improve their teaching; (c) significantly increased teachers' access to useful ideas for teaching; (d) enhanced the quality of learning of students; (e) had a positive effect on project teachers' teaching in non-project classes; and (f) beginning teachers felt supported not intimidated by this process.

Ingvarson (1990) also reported that the positive responses increased as teachers had more experience with moderation, which again reflects the need for time to become knowledgeable and skilled in using this process. This experience seems to hold true for at least one of the ADCs that had used the moderation process during the first year as well. This ADC's director reported that the teachers have changed. They have begun to appreciate the moderation process more. They used moderation to inform instruction and are now convinced of its instructional value. The director noted that in the first year some of the teachers were making up their own multiple choice tests, but that this year they "don't use the old assessments anymore."

In our preliminary analysis of the effect of moderation, we have encountered similar benefits for teachers using the SEPUP assessment system. We urge caution however in drawing conclusions from this early work on the impact of the assessment system. The ADC teachers were functioning in a "hand to mouth" existence with the assessment system. That is to say, we were passing along some pieces of the assessment as they were developed. Since we were revising the assessment to be embedded in the revised course, we were always one step behind the curriculum developers. The teachers particularly wanted the course blueprint, which of course, was not final until the course revisions were done. One teacher even said that "the challenge this year was in not knowing exactly where the assessment part was leading."

There are many benefits to be had by teachers engaging in local assessment moderation, but there are barriers and limitations that we also discovered as we evaluated the implementation of the SEPUP assessment system in the six Assessment Development Centers (ADCs). We discuss both the benefits of and the potential barriers to successful implementation below.

Successful Features of the Local Assessment Moderation Process

To provide a framework for discussing the extent to which the SEPUP local assessment moderation groups attained the purposes initially set forth, we summarize the findings from our preliminary analyses for each of the three stated purposes of the local assessment moderation model.

Quality assurance.

The teachers in the ADCs learned how to use the scoring guides to assess student performance on the embedded tasks. In the local assessment moderation process, the teachers reached consensus on students' scores, then selected exemplar papers from among those scored to represent as many of the score levels as possible.

We collected these exemplar papers from five of the six ADCs, then in July 1995 we worked with a team of ADC teachers and directors to review exemplars from all sites. The team selected a refined set of exemplars from the samples that had been turned in during the field test. This year (1995-96), the Kentucky ADC teachers have continued to work with us to field test the refined set of exemplars and to collect exemplars from activities for which none were received. These exemplars again will be chosen through local assessment moderation.

Consistency of scores was a desired outcome of the local assessment moderations. During moderation, teachers discussed the interpretation of words such as "options" or "some versus all" in the context of understanding what was meant by particular score levels. One teacher in the Oklahoma ADC summed this up rather well; she wrote: "No matter what you do you're going to have problems with semantics and how you interpret what the scoring guide wants. During our moderation meetings we come to a common consensus of what should be contained at each score level (how we will interpret the scoring guide) that keeps us uniform in our scoring." In addition to this anecdotal evidence, analyses of the consistency of scores (for one teacher over time, and for groups of teachers at the site) are currently underway, as part of the quantitative procedures for quality control (see Wilson and Draney, 1996).

Professional development.

As noted earlier, we believe the local assessment moderation process fulfills many of the basic tenets of high quality professional development for teachers (see for example: Fullan, 1982, 1990; Little, 1993; Little & McLaughlin, 1991; Loucks-Horsley et al., 1989; Shepard, 1995). In particular, the meetings provide teachers an opportunity to interact, share ideas, and help each other at the point when they need the most support, that is, when they are implementing an innovation (Fullan, 1982; Guskey, 1986; Ingvarson, 1990; Keiny, 1994).

Through local assessment moderation teachers at various stages of development in their understanding of this embedded assessment system receive sustained support to: (1) implement the innovation; (2) share with other colleagues engaged in the same project; (3) engage in inquiry about their instructional and assessment practices; (4) at times, argue over philosophical differences and yet reach consensus; and (5) reflect on their own practice and begin to modify their instruction based on their reconstruction of the role of assessment in teaching.

As a model of ongoing professional development and support, we observed the local assessment moderation to be invaluable. Indeed, we cannot see how teachers could engage in such a complex innovation as the SEPUP assessment system without such sustained support. The teachers were able to learn about assessment, use assessment in their classrooms, and then receive support from colleagues engaged in the same professional activity. The SEPUP teachers' comments paraphrased earlier echo many of the elements of high quality, successful professional development noted in the literature.

Learning about assessment. Most of the teachers participating in the SEPUP ADCs were able to develop an understanding of the assessment components. For example, most teachers reported that the moderation process helped them to understand the scoring guides and that over time they became easier to use. As Little and McLaughlin (1991) note, professional development is enhanced when teachers have a shared, common purpose, and such was the case with SEPUP assessment. The local assessment moderation provided time for teachers to reflect on their use of the assessment system and what this meant substantively in terms of changing instructional practices. For example, teachers reported that they were better able to keep track of students' growth over time on course variables.

Using assessment in the classroom. The teachers indicated that through moderation they were able to use the assessment tools with students more effectively. For example, they used the scoring guides to clarify expectations for student performance. Other teachers taught students how to moderate each others' papers, so that the students learned how to use the scoring guides as well.

Teachers grew more conscious over time to the fact that the assessment moderation process was guiding their instruction and enabling them to evaluate their own teaching as well as assess their students' learning. For example, teachers learned through moderation that their expectations for student performance were sometimes set too high or too low, and they made adjustments. Some teachers also shared other students' papers from the moderation sessions with their students to provide examples of the various scoring levels. Other teachers reported that they were able to recognize student deficiencies, and consequently covered concepts, such as ratios, so that their students would be prepared to move forward in their understanding of concepts like parts per million.

All of the teachers grappled with the issue of converting scores to grades to meet local requirements. In the beginning, there were many complaints about the time required to score the assessment activities or the need to "double score" (i.e., score once for the SEPUP score, then re-read the paper and assign a grade). Teachers learned over time that scoring, although still more time-intensive than scoring multiple choice tests, became easier as they internalized the scoring guides. The teachers became more adept at scoring as their experience grew, and learned to incorporate it into their local grading practices. In time, the complaints were replaced by comments about the value of the assessment information.

Providing collegiality and peer support. The SEPUP teachers were provided an opportunity to share and collaborate with colleagues as part of the development of the SEPUP course. Can teachers easily use the embedded assessments and scoring guides on their own? In our estimation, the professional growth for teachers comes from the interaction with colleagues and not in isolation using the assessment materials. Teachers described how they internalized the scoring guides and yet when they came together as a group for moderation, they saw how other teachers used the same scoring guides and this opportunity to reflect provided new insights into their own teaching as well as their students' understanding of the materials.

One teacher in the Alaska ADC described the early moderation sessions as "painful" and people left with "hurt feelings." She went on to say that "we all may have had different expectations in the beginning" but she was "happy with how it jelled." Others from this same center also indicated that having the time and space to work out "philosophical differences" was important to developing the group's rapport. Without this rapport, a group could sink into a perfunctory abyss as we found to be the case with another center that never "jelled" and in which the teachers remained very defensive about their students' scores throughout the year.

Feedback.

During the field test, all ADC teachers were asked to complete Assessment Feedback forms for each activity that they used and scored. In this way, we collected formative information on all activities done, not just those moderated. We used this information to revise the activities and to begin to design an assessment manual, that was also worked on by the ADC representatives in July 1995. We were looking for information such as the match between the prompt and the activity, or how congruent the scoring guide or variable selected for assessment was with the activity.

During local assessment moderation, the teachers completed Quick Writes about their initial reactions after scoring student papers. They were to report on what was helpful when using the scoring guide as well as the difficulties encountered. Further, they were asked to provide specifics on improving the scoring guides and the prompt or activity. Teachers were also asked how they would change the course or their instruction of the material before doing the assessment again. Finally, they were asked how they would apply this assessment information to further their students' progress on the variable assessed. In some cases SEPUP teachers felt so pressed for time that the responses to these questions were generally brief or left unanswered. Others provided more thorough responses that helped in revisions to the scoring guides, assessment activities and other assessment components at the conclusion of the field test . If an activity was particularly problematic, we did receive more extensive feedback. Comments such as "there is too great a jump between 0 and 1" helped us revise the score level definitions. In another example, ADCs reported struggling with the "integration" element for the *Evidence and Tradeoffs* variable. In the final revision, this scoring guide has two not three elements, having collapsed the integration concept into the *Using Evidence to Make Tradeoffs* element, while maintaining the *Using Evidence* element. The call for exemplar papers was loud and clear from all ADCs. Exemplars are not only considered valuable for supporting teacher's scoring, but also as a means to clearly set expectations for students. The most commonly reported instructional implication of the moderation process was the need to identify expectations for students as clearly and unambiguously as possible *before* the assessment activity was initiated.

Teachers were asked to complete a Moderation Reflection form at the end of each local assessment moderation. They were to reflect on their current understanding of the scoring guide or variable. Teachers were also asked if the moderation changed their ideas about how to modify the course activities preceding the assessment or how they would move forward in their instruction. As with the Quick Writes, we received varying responses on the Moderation Reflection forms. At times, teachers aired their frustrations while others said little or nothing. Other times teachers indicated a variety of benefits to the moderation process, including but not limited to: collegial support; clarification on scoring guides and variables; insight on how others were using scores for grading purposes; and ultimately as a place to become "recharged" about their involvement in the SEPUP field test. Collegial support manifested in several ways. One common example was that teachers who were further behind in the course activities were able to learn about difficulties that the other teachers had been encountering, so that they could avoid similar pitfalls. Another example was cited by a beginning teacher, he reported that he was able "to see how other teachers graded" during moderation and this was very helpful for someone new to the classroom.

Teachers also described what they learned from moderation in terms of their students' understanding and what they needed to do to improve their instruction. For example, early in the course an Oklahoma ADC teacher noted: "I need to start getting them to question the validity of their evidence. This is not a skill that I have required of them to date. They must develop it in order to go on to the next score level. I know it will develop with time." About three-quarters into the school year, this same teacher reported that: "I need to encourage my students to develop the ability to provide additional information and question the source of information." In the beginning, she was using the vocabulary of the scoring guide (e.g., "validity of evidence"), but by the end, she had reconstructed this scoring guide language into something that both she and her students could more easily understand. Since "questioning the source, validity, and/or quantity of evidence" is part of the description of the highest score level (4) for the element *Using Evidence to Make Tradeoffs*, it is also worth noting that this issue resurfaces this far into the school year. A level 4 response is highly unlikely early in the course on this element because students are grappling with the concepts of evidence and tradeoffs, and later build on this knowledge to weigh the evidence to make tradeoffs.

We also asked that the local assessment moderation sessions be tape recorded[4]. These tapes provide much richer insight into the value of the moderation process than all the forms put together. Sometimes the tapes were turned off midstream in order for comments to be made off-the-record, and other times ADCs forgot to tape a session or had mechanical difficulties. One can quickly *hear* the difference between a very successful moderation (rich dialogue, pointed comments, questions, and so on) and a less successful moderation (teachers announce scores perfunctorily and do not engage in any discussion of substance). Some of these moderation sessions were also observed by assessment team members, and the observations corroborate the evaluation of the level of success with moderation.

Contextual Factors that Affected Implementation

The level of success with implementation of the local assessment moderation was determined to a large extent by the organizational context, that is, the Center mattered. Features of the ADCs that mattered include: strong leadership; institutional support; and teacher proximity and collaboration.

Leadership for change.

The ADC teachers in general came from different schools, so their experiences with "norms of collegiality" (Little, 1982; McLaughlin, 1993) varied. As noted earlier, some felt isolated in their schools while others participated in science department or grade level teams. The key to bringing the group together cohesively was the leadership provided by the ADC director or from strong teachers with prior SEPUP experience in the group. The importance of leadership spans the gamut of ensuring that the group develops a rapport to preparing teachers to use local assessment moderation to facilitating moderation sessions.

Who the leader is may not be all that important, but according to the ADC teachers, there needs to be someone in this position to move the group toward consensus, to intervene in personal conflicts, to diffuse philosophical differences that digress from the work at hand, and overall to keep the group on task. Time, as always, is a factor for teachers, so maintaining task orientation is important.

Institutional support.

As Little (1993) and others have suggested, a quality professional development program needs to balance support for institutional initiatives with support for those initiated by teachers. For some of the ADCs, the purpose of changing teachers' assessment practices was consistent with state initiatives (such as the Kentucky Educational Reform Act) or local school or district-level staff development goals. These goals were balanced with the needs of the participating teachers, who for the most part were very successful science teachers, but were interested in learning about new assessment strategies. In terms of school-based

[4] Tapes of these sessions helped frame the exemplary moderation interchange provided earlier.

administration, teachers need support from their principals, but in the case of SEPUP this was generally not a problem. Most of the SEPUP teachers probably would have received support to do almost any reasonable thing they wanted because they are respected by their principals and have a long history of involvement in innovations.

Strong leadership was also important in securing district-level support. The most critical support factor related to the district was gaining access to the ADC teachers for whole-day or half-day meetings rather than after-school sessions, including release time[5] and substitutes. Having sufficient time as a group was important for three fundamental reasons, time is needed: (1) to build a group rapport or a SEPUP 'norm of collegiality,' to borrow Judith Warren Little's (1982) term; (2) to build teacher understanding of the SEPUP assessment system; and (3) for teachers to function as reflective practitioners. Having district support has broader implications as well, but with time being an oft-noted issue for teachers, it is critical to not lose sight of the opportunity cost of teachers' time.

In two of the three ADCs visited in May 1995 near the end of the field test, the ADC director was a District person. In the third case, the ADC director was a classroom teacher. This latter ADC was also unique in that all five teachers in the ADC were from different schools and different districts, which severely minimized district support for the project. Further, the ADC director admitted that his power to facilitate the group was minimal because of the lack of a common denominator, such as a district. Without firm institutional support for this group of teachers, their experience was less than successful.

Teacher proximity and collaboration.

Having teachers from more than one district can work; the Alaska ADC was a good example of this. Two teachers in one district north of Anchorage communicated with each other between moderation meetings, while the teachers in the Anchorage School District tended to call the ADC Director or one of the experienced SEPUP teachers. Given the long commute to attend meetings for three of the teachers, the Alaska ADC met for whole days. However as noted above, too many districts can be detrimental and minimize the leadership of the ADC director.

Teacher collaboration is related to proximity in that teachers who are closer tend to be able to spend more time communicating with one another. One of the Alaska ADC teachers noted that she would have liked to confer with one of the teachers from the other district, but that it was a long distance call, so she tended not to contact her between meetings. Teachers most often called one another to ask a question about an assessment task or to clarify something on a scoring guide. On occasion, teachers would actually meet or pursue a mutual project together.

In the three sites visited in May 1995, the quality of moderation varied, but so did the amount of experience with using local assessment moderation and the amount of time devoted to staff development. As noted above, the full days that Alaska used were critical not only to forming a cohesive group, but to learning about the moderation process and learning from each other how the assessments were working in their classrooms. In this way, the moderation meetings became the ongoing support that teachers needed as they implemented the SEPUP course.

[5] The ADCs received a budget for participation, but some chose to pay teachers' stipends and meet after school, while others paid for substitute costs with the district providing teacher release time.

Conclusions and Directions for Further Research

The teachers who participated in the ADCs have contributed volumes of information about the assessment system. In our efforts to honor their feedback, we have revised the assessment materials and the local assessment moderation process to incorporate the suggestions and best practices that the SEPUP teachers have shared with us.

In terms of the three purposes of assessment moderation, we have noted above that (a) the moderation process provided a critical forum for teachers to learn about the rating process, and to achieve local consensus, (b) it was instrumental in developing awareness, confidence and knowledge in the system itself, and in the potential role of assessment in instruction, and (c), it provided crucial formative information to the assessment developers.

According to the field test teachers and the ADC directors, teachers need ongoing staff development and support to use the SEPUP Assessment System. Although the terms "embedded assessment" and "rubrics" or "scoring guides" have been in existence for quite some time, teachers still need support in changing their beliefs and instructional behavior when it comes to assessment. There is a need for initial preparation to use the assessment system, including the embedded activities, the scoring guides and local assessment moderation.

Teachers do not need to be at the same school for the SEPUP Assessment System including moderation to work, although with fiscal constraints this might be the ideal. In Alaska and Kentucky, both with district-level change initiatives, the SEPUP teachers were all from different school sites. Most of the SEPUP teachers had some connection with an on-site teacher, either another science teacher or an interdisciplinary team member. The support from local colleagues was generally quite good.

As a framework for *further research* on the usefulness of local assessment moderation consider its three purposes, as laid out above. First, in enhancing technical quality, we need to observe the implications of the use of developmental maps in the moderation process (i.e., the maps are based on the same data as the above results concerning moderation, and hence were not available in the Field Test year). Detailed work will be needed to track and understand the influence of differing moderation groups on the technical consistency with which the ratings are carried out. In the long run, we will want to consider the effects of the inevitable development of local "understandings" of the scoring guides and their use. Second, in supporting teachers' professional development, we will need to make a serious effort to relate our findings to that of other researchers in the area, such as Little and McLaughlin (1993, p.6) who describe three “largely unexplored dimensions” of teachers' professional interactions: intensity, inclusivity, and orientation. These are described in the context of collegial relations within schools, but can be extrapolated and applied to other groupings of teachers, such as the ADCs. We will not attempt to do so here, but note that these dimensions have counterparts in the moderation experience. Third, with respect to improving feedback to developers, we have hardly begun to scratch the surface of issues for systematic study. Issues in this area are discussed in the accompanying paper by Wilson, Thier and Sloane (1996).

Reference List

Abbott, J. (1991). *Primary assessment program writing moderation booklets: A report of the survey of their usage and usefulness*. (Northern Territory Department of Education, Darwin, Australia). (ERIC Document Reproduction Service No. 351 373)

Baker, E.L., O'Neil, H.F., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist, 48* (12), pp. 1210-1218.

California Department of Education. (1995). *A sampler of science assessment.* Sacramento, CA: Author.

Chittenden, E. (1991). Authentic assessment, evaluation, and documentation of student performance. In V. Perrone (Ed.), *Expanding student assessment* (pp. 22-31). Association for Supervision and Curriculum Development.

Fullan, M. (1982). Professional preparation and professional development. In *The meaning of educational change* (pp. 257-287) New York: Teachers College Press.

Fullan, M.G. (1990). Change processes in secondary schools: Toward a more fundamental agenda. In M.W. McLaughlin, J.E. Talbert, & N. Bascia (Eds.), *The contexts of teaching in secondary schools: Teachers' realities* (pp. 224-255). 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, (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.

Harlen, W. (1994). *Concepts of quality in student assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. 367 712)

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)

Keiny, S. (1994). Constructivism and teachers' professional development. *Teaching and Teacher Education, 10* (2), 157-167.

Linn, R.L.. (1993a). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15* (1), pp. 1-16.

Linn, R.L.. (1993b). Linking results of distinct assessments. *Applied Measurement in Education, 6* (1), pp. 83-102.

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

BEST COPY AVAILABLE

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), pp. 15-21.

Little, J.W. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal, 19* (3), pp. 325-340.

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

Little, J.W. & McLaughlin, M.W. (1991). *Urban Math Collaboratives: As the teachers tell it.* Unpublished Manuscript. Center for Research on the Context of Secondary School Teaching, Stanford University.

Little, J.W., & McLaughlin, M.W. (1993). Introduction: Perspectives on cultures and contexts of teaching. In J.W. Little & M.W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, and contexts* (pp. 1-8). New York: Teachers College Press.

Loucks-Horsley, S., Carlson, M.O., Brink, L.H., Horwitz, P., Marsh, D.D., Pratt, H., Roy, K.R., & Worth, K. (1989). *Developing and supporting teachers for elementary school science education.* Washington, D.C.: The National Center for Improving Science Education.

Loucks-Horsley, S., & Stiegelbauer, S. (1991). Using knowledge of change to guide staff development. In A. Lieberman & L. Miller (Eds.), *Staff development for education in the '90s: New demands, new realities, new perspectives* (2nd ed.; pp. 15-36). New York: Teachers College Press.

McCallum, B., Gipps, C., McAlister, S., & Brown, M. (1995). National Curriculum assessment: Emerging models of teacher assessment in the classroom. In H. Torrance (Ed), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment.* (pp. 457-87). Philadelphia, PA: Open University Press.

McLaughlin, M.W. (1991). Enabling professional development: What have we learned? In A. Lieberman & L. Miller (Eds.), *Staff development for education in the '90s: New demands, new realities, new perspectives* (2nd ed.; pp. 61-82). New York: Teachers College Press.

McLaughlin, M.W. (1993). What matters most in teachers' workplace context? In J.W. Little & M.W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, and contexts* (pp. 79-103). New York: Teachers College Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), pp. 13-23.

Messick, S. (1989). Validity. In Linn, R.L. (Ed.), *Educational Measurement* (3rd edition; pp.13-103). New York: Macmillan Publishing Company.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher, 23* (2), pp. 5-12.

New Zealand Qualifications Authority. (1992). *Designing a moderation system. Developing a qualifications framework for New Zealand.* (ERIC Document Reproduction Service No. 354 330)

Resnick, L. & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford and M.C. O'Connor (Eds)., *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-76). Boston, MA: Kluwer Academic Publishers.

Richards, A.L. (1992). *Moderation procedures in English and mathematics in years 10 and 11 in the northern territory* (Northern Territory Department of Education, Darwin, Australia). (ERIC Document Reproduction Service No. 351 372)

Ross, J.A. & Regan, E.M. (1993). Sharing professional experience: Its impact on professional development. *Teaching and Teacher Education, 9* (1), pp. 91-106.

Shepard, L.A. (1995). Using assessment to improve learning. *Educational Leadership, 52* (5), pp. 38-43.

Sloane, K., Wilson, M. & Samson, S. (1996). *Designing an embedded assessment system: From principles to practice.* Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Torrance, H. (Ed) (1995a). *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment.*. Philadelphia, PA: Open University Press.

Torrance, H. (1995b). Teacher involvement in new approaches to assessment. In H. Torrance (Ed), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment.* (pp. 44-56). Philadelphia, PA: Open University 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. (1992a). *The integration of school-based assessments into a state-wide assessment system: Historical perspectives and contemporary issues.* Unpublished paper: University of California at Berkeley.

Wilson, M. (1992b). Educational leverage from a political necessity: Implications of new perspectives on student assessment for Chapter 1 evaluation. *Educational Evaluation and Policy Analysis, 14*(2), 123-144.

Wilson, M, & Draney, K. (1996). *Mapping student progress with embedded assessments.* Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Wilson, M., Thier, H., & Sloane, K. (1996). *What have we learned from developing an embedded assessment system?* Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Wilson, M, & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement.*

Zessoules, R. & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perrone (Ed.), *Expanding student assessment* (pp. 47-71). Association for Supervision and Curriculum Development.

Appendix A:
SEPUP Teacher Moderation Process Used for ADC Training Summer 1994

Pre-Moderation: Teachers are to score student work and bring a class set of papers (i.e., 5 or 6 papers from one class) to the moderation, having chosen a range of example papers to represent as many score levels as possible and also difficult to score papers.

Part 1: Initial Reactions to Scoring Student Work (5 minutes)--Teachers complete the top half of a Quick Write form in order to focus on issues of scoring and assessment.

Part 2: Scoring Guide Discussion (15 minutes)--review assessment task and questions; review scoring guide(s) used for assessment; discuss scoring guide questions and concerns.

Part 3: Pair Discussion and Selected Rescoring of Student Work (45 minutes)--
Teachers are to select a partner and switch papers; each teacher "blindly" rescores the partner's papers. Teacher pairs come to consensus on a final score for each student paper and record this on a score report form.

Part 4: Final Group Identification of Exemplar Student Work for each Score Level (15 minutes)--select exemplar papers for each level and reach consensus on a set of exemplar papers (to be turned into the Assessment Project).

Part 5: Instructional Implications (20 minutes)--Moderation leader facilitates a discussion about instructional implications to encourage teachers to discuss how to improve their instruction, to discuss student work, and to make connections between the scoring guides and grading practices.

Part 6: General Discussion (15 minutes)--Teachers debrief on SEPUP, moderation, instructional, or management issues.

Part 7: Resolving Issues, Finding Solutions (5-10 minutes)--Teachers complete the bottom half of the Quick Write form to focus on issues of scoring and assessment changes due to moderation discussion and sharing student work.

Post Moderation: Student papers used for moderation, score report forms, altered scoring guides with exemplars, and instructional implications are supposed to be forwarded to SEPUP Assessment Project.

TM 025819

AERA April 8-12, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Local Assessment Moderation in SEPUP
[SEPUP: Science Education for Public Understanding Program]

Author(s): Lily Roberts, Kathryn Sloane, & Mark Wilson

Corporate Source:

Publication Date: March 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ ← **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
_____ Sample _____
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document → ☐

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
_____ Sample _____
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Lily Roberts	Position: Ph.D. Candidate
Printed Name: Lily Roberts	Organization: U.C. Berkeley
Address: 2344 Tiffany Way Chico, CA 95926	Telephone Number: (916) 894-3916
	Date: 4-18-96

ERIC
Full Text Provided by ERIC

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 27, 1996

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (23)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1996/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://tikkun.ed.asu.edu/aera/). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an AERA chair or discussant, please save this form for future use.

ERIC Full Text Provided by ERIC