

**Designing an embedded assessment system:
From principles to practice.**

Kathryn Sloane
University of Illinois (Champagne-Urbana)

and

Mark Wilson and Sara Samson
University of California, Berkeley

Abstract

In this paper, we discuss the principles guiding the design of the SEPUP Assessment System, describe the components of the full assessment system, and provide examples of each component to illustrate the ways in which the principles were realized in classroom and teacher tasks. We present four principles of an embedded assessment system, and discuss how they are to be developed through three aspects of development and implementation of the assessment system. We then describe the specific components of the assessment system that have been developed in SEPUP to realize these principles: a) the overarching framework of SEPUP "variables", which are the key concepts and skills that define the course, are reinforced throughout the year-long curriculum, and form the basis of the assessment system; b) assessment "blueprints" that guide the focus and placement of the assessment tasks; c) the assessment tasks and scoring guides, linked to specific variables; d) statistical and quality control procedures used to validate scoring procedures and produce measures of student performance on each of the variables; and e) feedback mechanisms, including variable maps that define student progress and performance over the course of the year. This paper provides motivation and an overview of the assessment system and the context for subsequent papers that focus on specific aspects of the system.

Keywords: embedded assessment, alternative assessment, science education

Acknowledgments

The project described here began in 1993 and has benefited from the contribution of all the members of the SEPUP Development Team (at the Lawrence Hall of Science) and the SEPUP Assessment Team (at the Graduate School of Education, UC Berkeley). In particular, we would like to acknowledge the work of Herb Thier, Barbara Nagle, Mike Reeske, and Bob Horvat, in the development of the IEY curriculum and in working with us to develop the embedded assessment materials. Members of the assessment team have also included Lily Roberts, Robin Henke, Amy Jackson, Megan Martin, Chris White, Harry Case, Eric Crane, and Karen Draney--all of whom made substantial contributions to the development and analyses of the assessment materials. And, we owe a special thanks to all of the teachers and administrators who participated in the field test of these materials (especially to Dick Duquin, whose efforts went well beyond) and who provided the valuable feedback for improvements to the system and its implementation.

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 author and do not necessarily reflect those of the granting agencies.

Introduction

In recent years, "alternative assessment" has become a major topic of interest, debate, and experimentation in the nationwide efforts at educational reform. Initial hopes that "alternative", "authentic", or "performance" assessments of student achievement would drive (or at least facilitate) changes in what and how students are taught have been tempered by the realities of implementation. Efforts to introduce alternative assessments into large-scale, high-stakes state and district testing programs have met with mixed results due to high costs, logistical barriers, and political ramifications (e.g., Gipps, 1995; Rothman, 1995). For example, the demise of the CLAS testing system was due principally to the complications of using performance assessments, both political and financial. Efforts to introduce alternative assessments into ongoing classroom practices have been less publicized, but have also met with issues relating to costs (primarily in terms of time) and teachers' level of preparation and acceptance (e.g., Chittenden, 1991; McCallum et al, 1995; Shepard, 1995). The rationale for developing and using alternative assessment is quite compelling, however. Alternative assessments offer the potential for greater "ecological validity" and relevance, assessment of a wider range of skills and knowledge, and adaptability to a variety of response modes, compared to traditional tests (e.g., Baron, 1995; Gardner, 1992; Malcom, 1995; Wiggins, 1989, 1993).

Over the past few years, a growing number of states have experimented with alternative forms of assessment as part of their educational reform packages (Rothman, 1995). These new forms of assessment include attempts to use a variety of types of items, to emphasize performance and understanding, and to link the new assessment procedures to new state frameworks or guidelines. Much has been learned, but much still needs to be learned, about the costs, technical issues, teacher development, and public reaction to these new forms of assessment (Worthen, 1993).

Although these "high stakes" alternative assessment programs are related to frameworks or curricular standards, they are not usually linked to a specific curriculum or instructional context. Therefore, the "match" between what is actually taught and what is eventually assessed can still be an issue. But for assessments to be truly "authentic", they must be linked to a specific instructional

context and, ideally, occur as part of the ongoing teaching and learning process---and that occurs within the classroom. If alternative assessment practices are to have their greatest impact on instruction, they need to be seen as integral to (i.e., necessary to) good instruction, rather than as “add-ons” or “external demands.”

The need to *integrate* assessment into the curriculum and instruction process (i.e., the classroom context) has been emphasized by a number of researchers (e.g., Brown, et al, 1992; Glaser, 1987; Resnick and Resnick, 1992). Explanations of the distinctions between alternative assessment procedures and more traditional forms of testing frequently emphasize the importance of teachers' roles in mediating and interpreting the alternative assessment results within the classroom context (e.g., Chittenden, 1991; Wolf, et al, 1991), as well as the more immediate and meaningful uses of the alternative assessment procedures in the ongoing instructional process (Cole, 1991).

But it is also recognized that such integration will require new views of the teaching and learning process, new roles for (and demands on) teachers, or even a new “assessment culture” in the classroom (Brown, et al, 1992; Cole, 1991; Resnick and Resnick, 1992; Torrance, 1995a; Zessoules & Gardner, 1991). Preparing teachers to use these types of assessments in their classroom teaching may be a difficult challenge. Teachers' understanding and acceptance of innovations are crucial to the ultimate success of change (Airasian, 1988; Stake, 1991). But simply introducing more “performance-type” measures into classroom assessment will not necessarily produce immediate changes in teachers' conceptualization of assessment, in the form of instruction, or in the use of assessments in teachers' instructional planning and decision-making (McCallum, et al, 1995; Shepard, 1995; Torrance, 1995b).

For classroom-based alternative assessment procedures to gain “currency” in the assessment community, issues of technical quality will have to be addressed, as well. Despite the plea of Wolfe et al (1991), the development of practical procedures for establishing the technical quality of classroom-based alternative assessments lags behind that for high stakes assessment programs.

There *have* been some developments in thinking about the way that formal assessment systems can integrate information from a variety of sources. One such is called an *assessment net* (Wilson & Adams, 1996), which is composed of: (a) A framework for describing and reporting the level of

student performance along achievement continua; (b) the gathering of information through the use of diverse indicators based on observational practices that are consistent both with the educational variables to be measured, and with the context in which that measurement is to take place; and (c) a measurement model that provides the opportunity for appropriate forms of quality control. The assessment net concept is the basis for the formal measurement approach used below.

In this paper, we describe the development of a classroom-based assessment system which builds upon methodological advances in alternative assessment techniques and attempts to address salient issues in the integration of alternative assessment into the classroom teaching and learning context. The assessment system was developed for, and is integrally connected to, a specific curriculum in issues-oriented science for the middle grades. We describe the principles that guided the creation of the assessment system, the component parts of the system and how they work together, and some of the lessons learned in implementing the system in school and classroom contexts. The system we developed and put into place offers one model of how assessment can be incorporated into the classroom teaching and learning process. As such, it provides a "case" of sorts on how theoretical principles of assessment can be translated into practice, and some of the problems and pitfalls encountered.

Principles of an Embedded Assessment System

In designing the assessment system for *Issues, Evidence, and You*, we were guided by four "principles" of assessment. The principles represented the standards or ideals that should, we believed, be reflected in a technically-sound, curriculum-embedded, classroom-based system of student assessment. The roots, or foundation, of these principles can be traced (in part) to recent work in measurement theory and the recent research literature on alternative assessment practices. However, the combination of these principles, and the interrelationships among them, represent a new approach to classroom assessment.

1. Developmental Perspective:

The first principle is that an assessment system should be based on a developmental perspective of student learning. Assessing the development of students' understanding of particular concepts and skills requires a model of how student learning develops over a set period of (instructional) time. A developmental perspective helps us move away from "one shot" testing situations, and away from cross sectional approaches to defining student performance--toward an approach that focuses on the process of learning and on an individual's progress through that process. Clear definitions of what students are expected to learn, and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material, are necessary to establish the construct validity of the assessment system.

2. Instructional Fidelity .

The second principle focuses on the "match" between what is taught and what is assessed. This principle represents, of course, a basic tenet of content validity: that the items on a test are sampled appropriately from a domain that is defined by the content and the level of cognitive processing expected in a given body of instruction. Traditional testing practices--in "high stakes" or standardized tests as well as in teacher-made tests--have long been criticized for over-sampling items that assess only basic levels of knowledge of content topics and ignore more complex levels of understanding. The rationale for the development of "authentic" or "alternative" or "performance" assessment techniques is based, at its heart, on the need for a better match between important learning objectives (e.g., "problem solving") and the methods by which student performance on these objectives is assessed. As more attention is directed toward changing curricular materials and instructional methods--to reflect constructivist theories of learning and/or to reflect "higher-order" learning objectives--the mismatch between curriculum, instruction, and assessment can become even more pronounced.

Concerns about the match between curriculum, instruction, and assessment are discussed from both the curriculum development and assessment perspectives. From the curriculum perspective, efforts to emphasize new approaches to teaching and learning are inhibited by the form and content of accountability tests. Reports abound of teachers interrupting their use of new curricular materials in

order to "teach the material" that students will encounter on the district- or state-wide tests. From an assessment perspective, advocates of assessment-driven reform hope to take advantage of the tendency to "teach to the test" by aligning "high-stakes" testing procedures to the goals of curricular reform. 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" (p. 59).

3. Teacher Management and Responsibility

The third principle, therefore, that must be considered in building a classroom based assessment system is that teachers must be the managers of the system, and hence must have the tools to use it efficiently and use the assessment data effectively and appropriately. New forms of assessment, while perhaps more valid and more interesting and challenging to the student, make new demands on the teacher. For example, how can a teacher focus on rating one student's "performance" to the exclusion of monitoring other students' activities? How can teachers manage the additional time demands of scoring open-ended responses generated by 150 students? How can qualitative statements describing levels of performance be translated into letter grades, as required (and expected) by administrators, parents, and the students themselves? Any successful classroom based system must take into account the demands placed on the teacher for administering, scoring, interpreting, and reporting student performance.

There are two broad issues involved in the Teacher management and Responsibility principle. First, it is the teachers who will use the assessment information to inform and guide the teaching and learning process. Alternative assessments conducted as part of district or statewide accountability programs, no matter how valid or appropriate to what is taught in the classroom, cannot provide the immediate feedback to teachers that is necessary for instructional management and monitoring (Haney, 1991; Resnick & Resnick, 1992). For this function of assessment:

- (a) teachers have to be involved in the process of collecting and selecting student work
- (b) teachers must be able to score and use the results immediately--not wait for scores to be returned several months later, and

(c) teachers must be able to interpret the results in instructional terms.

Only then will teachers really be able to use the assessment system.

Second, issues of teacher professionalism and teacher accountability demand that teachers play a more central and active role in collecting and interpreting evidence of student progress and performance (Tucker, 1991). If they are to be held accountable for their students' performance, teachers need a good understanding of what students are expected to learn *and* of what counts as adequate evidence of student learning. They are then in a better position, and a more central and responsible position, for presenting, explaining, and defending their students' performances and the "outcomes" of their instruction.

4. *Quality of Evidence*

It is not sufficient that alternative forms of assessment should express new ideas of validity, they must also maintain the standards of fairness (such as consistency and unbiasedness) that have been accepted as standards for traditional assessments. Doing so involves many qualitative and technical challenges. On a logistical level, using "open-ended" or "performance-based" tasks require different procedures for collecting, managing, and scoring student work. Records of performances must be catalogued and stored. Responses can no longer be scanned by machine and entered directly into a statistical data base. Raters must score the work, and this raises issues of time and cost, as well as technical issues involved in rater fairness (e.g., consistency and reliability). There has been a tendency for the arguments surrounding "new" and "conventional" forms of assessment to be framed as a shift from an emphasis on reliability to a stronger focus on validity. This argument is bolstered, perhaps, by the long-accepted truism that teacher-made (i.e., "classroom-based") student assessments have greater "curricular" or instructional validity, in some sense, but will not have the strong technical properties of more carefully constructed standardized tests.

For classroom-based assessment to gain currency in educational reform, we contend that these assessments must be held to standards of fairness in terms of quality control. Teachers will continue to construct "teacher-made tests" and will rarely take the steps to establish the comparability or validity of these instruments. However, classroom based assessment procedures can be developed for specific

curricula and made available to teachers' use and adaptation. The evidence generated in the assessment process should be judged by its suitability for purposes of individual assessment and for purposes of evaluating student performance, instructional outcomes, or program effectiveness.

To ensure comparability, procedures are needed (i) to examine the coherence of information gathered using different formats, (ii) to map student performances onto the achievement variables, (iii) to describe the structural elements of the accountability system—tasks and raters—in terms of the achievement variables, and (iv) to establish uniform levels of system functioning. The traditional elements of test standardization, such as validity and reliability studies, and bias and equity studies, must be carried out within the quality control procedure. To meet this need we propose the use of generalized item response models (sometimes called item response theory). Generalized item response models such as those described by Adams and Wilson (1992), Kelderman (1989), Linacre (1989), and Thissen and Steinberg (1986), have now reached levels of development that make their application to many forms of alternative assessment feasible. The output from these models can be used as quality control information and can be used to obtain student and school locations on the achievement variables, which may be interpreted both quantitatively and substantively.

Aspects of Assessment Development, Implementation, and Use

The principles of assessment outlined in the preceding sections may be manifested in slightly different ways, or have different implications, depending on different "aspects" of the assessment process. We define three aspects:

Design and development: Establishing the goals, purposes, or "intent" of the assessment system, and defining the ways in which individual components of the system are conceived of and designed to work individually and in concert to achieve the intended goals and purposes.

Implementation: The translation of the designed components into actual field-based (classroom) practice; how the system works on a day-to-day basis.

Products: How the products that emerge from the implementation of the system are used by teachers, students, administrators, course developers, etc., as part of the educational (teaching and learning) process.

In Table 1, we display the principles of assessment development (columns) and the aspects of the assessment process (rows) in a matrix format. The cells of the matrix therefore define the implications of the principles for each aspect of the assessment process.

=====
Insert Table 1 about here
=====

For example, looking across the first row of the Table, in the Design and Development aspect of assessment, the principle of the developmental perspective dictates that a framework of variables be established to define what students are expected to learn and how this learning is expected to develop. At the same time, the Instructional Fidelity principle implies that assessments should be developed along with the instructional materials to ensure they are congruent with instruction in both form and content. The Teacher Management and Responsibility principle implies that the framework and the assessment tasks themselves should serve to reinforce and enhance teachers' understandings of the course goals (what students are expected to learn and how) as well as help them explicitly link student performance to the attainment of those goals. Finally, the Quality of Evidence principle implies that the design of the assessment tasks should conform to principles of sound technical design, such as the inclusion of discrete (defined) assessment opportunities and clear procedures for scoring student work. Thus, there are relationships between the principles, within each aspect of the assessment process.

The matrix also demonstrates the interrelationships among the aspects of the assessment process, because each of the principles provides a unifying "thread" throughout the process. For example, the Developmental Perspective principle implies that there should be an initial framework of developmental variables in the design aspect. This framework defines not only the content of student learning, but also the way in which student learning develops over time. The implication for the Implementation aspect is that each assessment, therefore, has a designated place in the instructional flow, reflecting the type of learning that students are expected to demonstrate at that point in time. In the Product aspect, scores assigned to student work can then be linked back to the developmental

framework, and used both to diagnose an individual's progress with respect to a given variable but also to "map" student learning over time (the "Products" aspect).

Adherence to each of the principles across each of the aspects of the assessment process produces a coherence or "internal consistency" to the system. Adherence to each of the principles within each aspect of the assessment process produces a comprehensive, or well-integrated system that can address the complexity of the classroom context and the desired linkages among curriculum, instruction, and assessment.

The matrix of assessment principles and aspects in the assessment process represents an ideal--standards to which a classroom based assessment system might aspire. In the SEPUP project, we have attempted to translate these ideals into a working assessment system linked to a specific curriculum. We turn next to a description of the SEPUP assessment system and the component parts developed for *Issues, Evidence, and You*.

The SEPUP Assessment System

The Science Education for Public Understanding Project (SEPUP) at the Lawrence Hall of Science has developed a year-long issues-oriented science course for the middle school and junior high grades entitled *Issues, Evidence, and You*. The course focuses on environmentally- and socially-contextualized science content. Societal Decision Making is a central focus of *Issues, Evidence, and You* and, in many ways, distinguishes this course from other middle school science courses. The goal of issue-oriented science is development of an understanding of the science and problem-solving approaches related to social issues *without* promoting an advocacy position. The concepts and skills needed to understand the process of Societal Decision Making form the basis of the SEPUP curriculum. As part of the course, students are regularly required to recognize scientific evidence and weigh it against other community concerns with the goal of making informed choices about relevant contemporary issues or problems.

An innovative and integral part of the SEPUP *Issues, Evidence, and You* course is its instructionally-embedded assessment and evaluation system which has been developed and implemented in concert with the SEPUP curriculum. This SEPUP Assessment System has been designed to apply new theories and methodologies in the field of assessment to the practice of teacher-managed, classroom-based assessment of student performance. The goal is not only to improve assessment methodologies but also to reform teacher practice in the use of assessment so it can be used immediately and reflectively to improve instruction and student achievement.

The SEPUP Assessment System is a comprehensive, integrated system for assessing, interpreting, and monitoring student performance in *Issues, Evidence, and You*. The full SEPUP Assessment System provides a set of tools for teachers to use to:

- Assess student performance on central concepts and skills in the curriculum;
- Set standards of student performance;
- Track student progress over the year on the central concepts; and
- Provide feedback (to themselves, to students, to administrators, parents, or other audiences) on student progress and on the effectiveness of the instructional materials and the classroom instruction.

The approach used is that of *embedded assessment*. Opportunities to assess student progress and performance are integrated into the instructional materials and are virtually indistinguishable from the day-to-day classroom activities. The student assessments were developed and field-tested along with the instructional materials and are considered by SEPUP to be an integral part of the teaching and learning process.

Initially, managing a new classroom-based system of embedded assessment demands much of the teacher. However, SEPUP believes that teachers must be recognized as the "front-line" professionals who ultimately determine the usefulness or effectiveness of any educational product or innovation. SEPUP developed this Assessment System to provide effective and efficient tools for teachers to use in collecting, interpreting, and presenting their *own* professional evidence on their students' achievements in *Issues, Evidence, and You*.

Components of the Assessment System

The Assessment System is built upon a foundation of three basic components:

1. **SEPUP Variables** - core principles, ideas, or topics focused upon in *IEY* which form the framework of the instruction and the assessment;
2. **Assessment tasks** - linked to the SEPUP variables and embedded in the instructional materials; and
3. **Scoring Guides** - criteria for assessing levels of student performance and interpreting student work.

Other parts of the System are designed to elaborate and to integrate these basic components. In addition to the basic components, the full System includes:

4. **Assessment Blueprints** - charts showing overview of all course activities, indicating SEPUP Variables and where assessment tasks are suggested;
5. **Exemplars** - samples of actual student work illustrating performance at each score level for all *Issues, Evidence and You* assessment tasks;
6. **Assessment Moderation** - process through which groups of teachers reach consensus on standards of student performance and discuss implications of assessment results for subsequent classroom instruction;
7. **Maps** - graphic displays used to record the progress of each student on particular SEPUP Variables over the course of the year; and
8. **Link Tests** - open-ended tests designed to repeatedly collect information on students' performance for the SEPUP Variables.

In the following sections, we describe each of these eight components in more detail, provide examples from the IEY assessment system, and discuss the relationship between the component and the principles of assessment we have proposed.

(1) SEPUP Variables

If we follow the Developmental Perspective principle, then we need to devise a framework of developmental variables that embody the learning that we expect students to experience in the IEY year. Hence, the *IEY instructional* materials and assessments are built around a core set of scientific concepts, processes, and skills that are central to the course. These concepts, processes, and skills are called "variables", and they form the framework for the entire course. All instructional objectives for each activity and all of the assessment tasks are linked to one (or more) of the five SEPUP variables.

The five SEPUP Variables are organized into three areas general categories:

A. Scientific Processes

Designing and Conducting Investigations (DCI) - designing a scientific experiment, carrying through a complete scientific investigation, performing laboratory procedures to collect data, recording and organizing data, and analyzing and interpreting results of an experiment.

Evidence and Tradeoffs (ET) - identifying objective scientific evidence as well as evaluating the advantages and disadvantages of different possible solutions to a problem based on the available evidence.

B. Science Concepts

Understanding Concepts (UC) - understanding scientific concepts (such as properties and interactions of materials, energy, or thresholds) in order to apply the relevant scientific concepts to the solution of problems.

C. Interaction Skills

Communicating Scientific Information (CM) - organizing and presenting results in a way that is free of technical errors and effectively communicates with the chosen audience.

Group Interaction (GI) - developing skills in working with teammates to complete a task (such as a lab experiment) and in sharing the work of the activity.

The first three variables - Designing and Conducting Investigations (DCI), Evidence and Tradeoffs (ET), and Understanding Concepts (UC) - are primary variables and are assessed most frequently. Students' performance on Communicating Scientific Information (CM) can be assessed in

conjunction with almost any activity or assessment, depending on the teacher's interest in monitoring student progress on this variable. Opportunities in the course have been indicated for assessing students' skills in this area. The final variable, Group Interaction (GI), is based on the SEPUP 4-2-1 model of instruction and can also be assessed throughout. We have developed a Scoring Guide for Group Interaction, but have left it to the discretion of the teacher to determine when s/he assesses individual or group progress on this variable.

Each of the five SEPUP Variables is defined by a number of sub-variables which are called "elements." The SEPUP Elements define how each variable is operationalized in the course. For example, Designing and Conducting Investigations (DCI) includes four elements: (1) Defining the Problem, (2) Selecting and Performing Procedures, (3) Managing Data, and (4) Analyzing and Interpreting Data. Students can be assessed on one or several elements. The Scoring Guide for each SEPUP Variable defines the elements and provides performance criteria for each score level.

Clearly the SEPUP Variables correspond to the Design and Development aspect of the Developmental Perspective principle. In the *IEY curriculum*, students are expected to develop on all five variable over the course of the year. The variables define what students are expected to learn and how this learning is expected to develop. It is from this framework of SEPUP Variables that the assessment system evolves. It focuses and anchors the assessments for the course.

(2) Assessment opportunities

Following the Instructional Fidelity principle, we need to create assessments that are an integral part of the instruction in IEY. Indeed, a variety of tasks are used for assessment in *IEY*. These include individual and group "Challenges," data processing questions, and questions following student readings. All assessment prompts are open-ended, requiring students to fully explain their responses. For each of the five SEPUP variables that define the framework of the course, opportunities to assess student performance are identified for teachers throughout the instructional activities. There are two types of assessment opportunities.

Assessments: which are points where a formal assessment of individual students' work are recommend. Assessments are linked to the SEPUP Variables (and to the relevant elements of

that variable that are being assessed) and are distributed throughout the instructional materials at critical junctures in the course, in order to provide teachers with a year-long picture of student progress and performance.

Quickchecks: which are opportunities to check for student understanding on important concepts, skills, or procedures that are being introduced or practiced in the activity, such as constructing a data table or preparing an oral report. Teachers may choose to check all students' work on these items, or only sample student work to get a sense of the class as a whole. Quickchecks may be scored "right/wrong" or by using the more detailed Scoring Guides.

As an example of an assessment prompt, the question shown in Figure 1 is taken from *IEY* Activity 19: "Is Neutralization the Solution to Pollution?" It is typical in that it requires students to integrate information from readings they did in previous activities' and labs in order to successfully answer the question.

=====
Insert Figure 1 about here
=====

As with most SEPUP assessments, this prompt has multiple components which must be considered, and it clearly outlines for the students what is expected in their responses. There is no one "right" nor "expected" answer; rather, students are required to make a statement or decision, and then justify it with the information and evidence they have learned through the activities. Their performance will be judged by the validity of the arguments they present, not simply the conclusion that they draw.

The Design aspect of the Instructional Fidelity Principle indicates the need to co-develop assessment and instruction (see Table 1.) The development of the SEPUP assessment tasks was done in conjunction with the course development (See Wilson, Thier and Sloane, 1996). This allowed them to be congruent with the instruction in both form and content.

The Implementation aspect of the Instructional Fidelity Principle describes the need for assessment to be integrated into regular instruction (see Table 1.) SEPUP assessments are embedded within the instruction of *IEY*. The prompts and tasks are indistinguishable in form or content from those used for instruction--they are indeed exactly those used for instruction.

(3) Scoring Guides

For the information from assessment opportunities to be useful to teachers, it must be couched in terms that are directly interpretable with respect to the instructional goals of the IEY variables. Moreover, this must be done in a way that is intellectually and practically efficient. Our response to these two issues are the SEPUP Scoring Guides. SEPUP Scoring Guides define the elements of each variable and describe the performance criteria, or characteristics, for each score level of the element. There is one Scoring Guide for each of the five SEPUP Variables, with each variable having between two and four elements. A student's level of performance on an assessment task is determined by using the Scoring Guide(s) for the variable(s) being assessed. The guide is used throughout the entire course for all assessments relating to a particular variable. This means that there will inevitably be a need for interpretation of the Scoring Guide for any particular assessment. We found that the combination of a uniform study guide with interpretation for individual assessments, was much more efficient for teachers than having independent Scoring Guides for each assessment.

Each SEPUP Scoring Guide uses the general structure shown in Figure 2. All SEPUP Scoring Guides share this structure but uses specific criteria to uniquely adapt them to individual SEPUP Variables and Elements. An example of the Evidence and Tradeoffs (ET) Variable Scoring Guide is found in Figure 3.

=====
Insert Figures 2 and 3 about here
=====

(4) Blueprints

In order to implement the Developmental Perspective principle, the teacher needs a tool that indicates when assessments might take place, and what variables they pertain to. The Assessment Blueprints are a valuable teacher tool for keeping track of when to assess students. Assessment tasks are distributed throughout the course at opportune points for checking and monitoring student performance. These points are indicated in the Assessment Blueprints. Teachers can review and plan for the progression of assessment tasks relating to each variable by using the blueprints.

The Assessment Blueprints, designed to correspond with the four parts of the *IEY* course, list all activities. Assessments and Quickchecks are indicated under the appropriate SEPUP Variables for each activity in which there is an assessment. In addition, the main *IEY* content concepts addressed by each Understanding Concepts assessment task have also been identified.

=====
Insert Figure 4 about here
=====

(5) Exemplars

In order to interpret each Scoring Guide, teachers need concrete examples (which we call “Exemplars” of the rating of student work. Exemplars, by providing concrete examples of what a teacher might expect from students at varying levels of development along each variable, represent a part of the Implementation of the Developmental Perspective Principle (see Table 1.) They are also a resource, available to the teachers as part of the Assessment System, which help them to be in charge of the system and hence are a part of the Implementation aspect of the Teacher Management & Responsibility Principle. (see Table 1)

Actual samples of student work, scored and moderated by teachers who pilot-tested the SEPUP Assessment System, are included with the SEPUP Assessment System. These illustrate typical responses for each score level for specific assessment activities. An example of a Level 1 response from Activity 5 - "John Snow and the Continued Search for Evidence" is illustrated in Figure 5.

=====
Insert Figure 5 about here
=====

Teachers have found the exemplars to be very helpful in learning to use the Scoring Guides since they provide concrete examples of student work at each scoring level. Exemplars are available for all SEPUP Variables except Group Interaction (GI) which have not been collected for logistical reasons. With practice, teachers may not need to refer to the exemplars; however, these are included as a resource for teachers' use whenever they find them helpful.

(6) Assessment moderation

The Teacher Management and Responsibility principle requires that teachers “take control” of essential parts of the assessment system, including the scoring process, and also demands that they grow professionally in order to master the system. We have devised the “Assessment moderation meeting” as part of our strategy to accomplish these goals. Assessment moderation also plays a crucial role in achieving the Quality of Evidence principle.

Moderation is the process in which teachers discuss student work and the scores they have given that work, making sure that scores are being interpreted in the same way by all teachers in the moderation group. In moderation sessions, teachers discuss the scoring, interpretation, and use of student work, and make decisions regarding standards of performance and methods for reliably judging student work related to those standards. Moderation sessions also provide the opportunity for teachers to discuss implications of the assessment for their instruction, for example, by discussing ways to address common mistakes or difficult concepts in their subsequent instruction.

The assessment moderation process translates the Implementation aspect of the Teacher Management & Responsibility Principle (see Table 1) into the classroom within the SEPUP system. The process gives teachers the responsibility of interpreting the scores given to students' work and allows them to set the standards for acceptable work.

The Products aspect of the Teacher Management and Responsibility Principle (see Table 1) is reflected in the way teachers use moderation to adapt their judgments to local conditions. Upon reaching consensus on the interpretations of score levels, teachers can then adjust their individual scores to better reflect the teacher-adapted standards

The sound technical design issue, or Evidence Principle, is translated into practice (see Table 1) by requiring teachers to make judgments about students' scores in a public way with respect to public standards. The use of moderation improves the fairness and consistency of the scores being given by various teachers.

(7) Variable Maps

To fully embody the Developmental Perspective principle (Products aspect) in the use and reporting of Assessment System results, we have developed “maps” of the SEPUP Variables. Variable maps are graphical representations of a variable, showing how it unfolds or evolves over the year in terms of student performance on assessment tasks. They are derived from empirical analyses of student data collected from *IEY* teachers' classrooms. They are based on an ordering of assessment tasks from relatively easy tasks to more difficult and complex ones.

Once constructed, maps can be used to record and track student progress and to illustrate the skills a student has mastered and those that the student is working on. A map of students' performance on the Evidence and Tradeoffs Variable can be found in Figure 6. By placing students' performance on the continuum defined by the map, teachers can demonstrate students' progress with respect to the goals and expectations of the course. The maps, therefore, are one tool to provide feedback on how students as a whole are progressing in the course. They are also a source of information to use in providing feedback to individual students on their own performances in the course.

=====
Insert Figure 6 about here
=====

Maps, as graphical representations of student performance on assessment tasks, can be used by teachers for their own planning and to show students, administrators, and parents how students are developing on the SEPUP Variables over the year. This Products aspect falls under the Developmental Perspective Principle (see Table 1.)

Maps, which are based on an ordering of assessment tasks from easy to more difficult on a particular variable, relate to specific parts of the *IEY* curriculum. This "curriculum-referenced" aspect of them illustrates how they fit into the Products aspect of the Instructional Fidelity Principle (see Table 1.)

As a result of teachers managing and using the SEPUP Assessment System, maps can be produced which allows them to assess both individual and class progress. This can then be used to inform instructional planning. For instance, if the class as a whole has not performed well on a variable following a series of assessments, s/he might feel the need to go back and re-address those

concepts or issues reflected by the assessments. This use of maps falls under the Teacher Management & Responsibility Principle (see Table 1.)

Both class and individual maps can make "scores" more meaningful. They can then be used (Use aspect) as the basis for evaluation by teachers and others. In this way, they operationalize the Evidence Principle (see Table 1.)

(8) Link Tests

In order to efficiently create the maps, we found that extra information was needed at regular points in the curriculum. Our response to this Quality of Evidence principle issue was to create "Link Tests" which are composed of short-answer items, each linked to a variable--an example is shown in Figure 7. The Link Tests also respond to the Teacher Management and Responsibility principle because they can be used as the basis for grading.

=====
Insert Figure 7 about here
=====

Link Tests are a series of tests given at major transition points in the *IEY* course. Each test contains open-ended items related to the content of the course which further assess students' abilities with the SEPUP Variables.

Some "Link Test" items are intentionally repeated throughout the year. The reason for this is so that teachers can measure student growth on the same (or related) items at the end of each Part of *IEY*. By re-assessing students on the same items, it is possible to see what additional insights they can provide as a result of further exposure to important concepts and ideas as the course progresses. Student scores on the Link Test provide another point of information to use in tracking student progress on the Variable Maps.

Items on the Link Tests can be used as an "item bank" for teachers to draw upon in designing their own end of unit or other tests to be administered during the year. Teachers can use the Link Test items as models of variable-linked, open-ended questions, or they may pull specific items to be included in other teacher-made tests.

Discussion

The account above of the principles of the Assessment System and their operationalization in SEPUP's *IEY* curriculum presents each component as though it were the results of a systematic planning process. This is not entirely accurate: Instead the principles were more like a wish-list we had when we started. As we proceeded to implement the various components of the system, the need for further components became clear, based mainly on feedback from teachers. The arguments presented above are mainly post-hoc rationalizations of the somewhat messy and disorderly steps to development.

In the papers that follow, we explore some specific issues that arose as we worked on the Assessment System: The maps that embody the Developmental Perspective; assessment moderation, which is so crucial to teacher development; and finally, we discuss some of the lessons we have learned as we designed and implemented the system. There are several other aspects of the Assessment System that also deserve to be addressed. Unfortunately there is not sufficient time in one symposium to address them also. These topics include: Evaluation of the success of the *IEY* curriculum with and without the Assessment System, in terms of student achievement and attitudes, and also teacher growth and satisfaction; technical aspects of the longitudinal analyses; and, what we have learned about performance assessments, both from the development process, and the empirical evidence. We will be working on these issues in the following months, and will report on them also.

References

- Adams, R.J., & Wilson, M. (1992). *A random coefficients multinomial logit: Generalizing Rasch models*. Paper presented at the annual meeting of the AERA, San Francisco.
- Airasian, P.W. (1988). Measurement-driven instruction: A closer look. *Educational Measurement: Issues and Practice*, 7(4), 6-11.
- Baron, J.B. (1991). Performance assessment: Blurring the edges of assessment, curriculum, and instruction. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 247-266). Washington, DC: American Association for the Advancement of Science.
- Biggs, J.B. & Collins, K.F. (1982). *Evaluating the quality of learning: The Solo Taxonomy*. New York: Academic Press.
- Brown, A.L., Campione, J.C., Webber, L.S., & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments* (pp. 121-212). Boston: Kluwer Academic Publishers.
- 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.
- Cole, N. (1991). The impact of science assessment on classroom practice. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 97-106). Washington, DC: American Association for the Advancement of Science.
- Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments* (pp. 77-120). Boston: Kluwer Academic Publishers.
- Glaser, R. (1987). The integration of instruction and testing: Implications from the study of human cognition. In D.C. Berliner & B.V. Rosenshine (Eds.), *Talks to teachers: A festschrift for NL Gage* (pp. 329-341). New York: Random House.
- Gipps, C. (1995). Reliability, validity, and manageability in large-scale performance assessment. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 105-123). Philadelphia: Open University Press.
- Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), *Expanding student assessment* (pp. 142-163). Association for Supervision and Curriculum Development.
- Kelderman, H. (1989). *Loglinear multidimensional IRT models for polytomously scored items*. Paper presented at the Fifth International Objective Measurement Workshop, Berkeley, CA.
- Linacre, J.M. (1989). *Many faceted Rasch measurement*. Doctoral Dissertation. University of Chicago.
- Malcom, S.M. (1991). Equity and excellence through authentic science assessment. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 313-330). Washington, DC: American Association for the Advancement of Science.

- 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* (pp. 88-104). Philadelphia: Open University Press.
- Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments* (pp. 37-76). Boston: Kluwer Academic Publishers.
- Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Francisco: Jossey-Bass Publishers.
- Shepard, L.A. (1995). Using assessment to improve learning. *Educational Leadership*, 52(5), 38-43.
- Stake, R. (1991). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan*, 71, 243-247.
- Torrance, H. (1995a). The role of assessment in educational reform. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 144-156). Philadelphia: Open University Press.
- Torrance, H. (1995b). Teacher involvement in new approaches to assessment. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 44-56). Philadelphia: Open University Press.
- Tucker, M. (1991). Why assessment is now issue number one. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 3-16). Washington, DC: American Association for the Advancement of Science.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models *Psychometrika*, 49, 501-519.
- Wilson, M., & Adams R.J. (in press). Evaluating progress with alternative assessments: A model for Chapter 1. In M.B. Kane (Ed.), *Implementing performance assessment: Promise, problems and challenges*. Hillsdale, NJ: Erlbaum.
- Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership*, 46(7), 41-47.
- Wiggins, G. (1993). Assessment: authenticity, context, and validity. *Phi Delta Kappan*, 75(3), 200-214.
- Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education*, 17, 31-74.
- Worthen, B.R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan*, 74(3), 444-457.
- 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.