

**Developing maps for student progress
in the BEAR Assessment System**

Mark Wilson and Karen Draney
University of California, Berkeley

January 1999

Key Words: Embedded Assessment, Science Assessment, Progress Maps.

Acknowledgments

This work would not be possible without the support of the SEPUP team at the Lawrence Hall of Science led by Herb Thier and Barbara Nagle. We would like to thank those individuals whose help was especially valuable in preparing the data for analysis, and in performing the numerous analyses which are summarized in this paper: In particular, we would like to thank Henry Stavisky, Chris White, and Eric Crane.

This research was supported in part by NSF grants Nos. MDR9252906 and ESI-9553548. Opinions expressed in the paper are those of the authors and do not necessarily reflect those of the granting agency.

Abstract

In this paper, we describe the development of psychometrically calibrated progress maps describing student performance throughout the year for a specific middle school science curriculum. The progress maps are based on an analysis of the structure of the curriculum in terms of “progress variables”, which summarize the important strands of student development that are intended by the curriculum developers. Each instructional activity in the curriculum is associated with at least one of these progress variables. The maps are calibrated using a combination of embedded performance tasks and link tests. The entire set of progress variables, maps and embedded tasks and link tests is an example of a generic assessment system—the BEAR Assessment System. We describe uses of the maps for individual students and whole classes, and also for evaluations. In terms of the latter, we show results that demonstrate that the assessment system itself can make a significant contribution to student learning.

Developing maps for student progress in the BEAR Assessment System

There are constant calls for sound evaluation of innovative curriculum projects in science (and every other school subject as well). The most common way to measure the success or failure of such curricula is through off-the-shelf standardized tests. These have a reputation for being “hard-nosed” and for offering an “even playing field” so that comparisons can be made across different curricula, two characteristics that we should certainly value in conducting evaluations. Unfortunately, in attaining these characteristics, standardized tests have also sacrificed two others: (a) consistency with the mode of instruction that we use in schools, and (b) congruency with the curriculum goals of our innovative curriculum projects. In contrast, the BEAR Assessment System¹ has been designed to maximize these latter two characteristics. We describe its broad outlines below, and describe in detail one of its principal features—*progress maps*.

The approach taken in the BEAR Assessment System is based on a number of recent developments in assessment. Following calls from many assessment specialists such as Wolfe (1991) and Stiggins (1991) for assessments that more validly measure what is being taught in classrooms, alternative forms of assessments found their way into many state testing programs. These have principally been in the form of performance assessments, although other types of assessments, including projects and portfolios, have also been adopted in some places. These general trends

¹ Named for the acronym of the Berkeley Evaluation and Assessment Research (BEAR) Center located in the Graduate School of Education at UC Berkeley, where this system was first developed.

have been reflected in science assessment with influential work being reported by Kulm and Malcolm (1991) and by Shavelson and Baxter (1992).

The BEAR Assessment System is built on the basis of embedded assessments that are principally in the shape of either performance assessments or projects. The prime determinant of the style of assessment is the purpose and content of the instruction in which the assessments are embedded, rather than any predisposition to work with only one type of assessment task. The basis for choice of assessment task is the style and content of the instructional events designed by the curriculum developers, which are then adapted for assessment purposes.

In this paper, we focus on one particular aspect of that system, the *student progress maps* which are used as the central connector between those aspects of the system which constitute “hard-nosed” measurement of student performance, and the interpretation of those results as instructional information by teachers. Before we can do that, however, it is necessary to provide an instructional context for the assessment system.

The SEPUP *Issues, Evidence, and You* curriculum

The Science Education for Public Understanding Program (SEPUP) at the Lawrence Hall of Science has developed a year-long issues-oriented middle school science curriculum entitled *Issues, Evidence, and You*. The course focuses on environmentally- and socially-contextualized science content. Societal Decision Making is a central focus of IEY and, in many ways, that focus distinguishes this course from other middle school science courses. The goal of issue-oriented science

is development of an understanding of the science content, the scientific methods, and the 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 IEY course is its instructionally-embedded assessment system which has been developed and implemented in concert with the SEPUP curriculum through a collaboration between SEPUP and the BEAR Center. This initial example of the BEAR 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 (see Sloane & Wilson (under review) and Wilson & Sloane (under review) for a broader discussion). 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.

Embedded assessment refers to the way in which IEY assessment tasks fit into the IEY curriculum. Traditionally, assessment activities are seen as separate from, almost an interruption to, instruction: Teachers give a series of instructional activities, then stop and administer an assessment, then continue with more instruction. In the assessment system that has been developed for IEY, assessment tasks are part of the regular instructional activities. Opportunities to assess student

progress and performance are integrated into the instructional materials and are indistinguishable from day-to-day classroom activities. Teaching and learning activities occur, and some of these activities are scored as formal assessments. Since these assessments are an integral part of the teaching and learning process, the assessments are intended to measure exactly what students are learning. The student assessments were developed and field-tested along with the instructional materials. A recent review by NSF staff of prominent middle school science curricula (NSF, 1997) found the IEY curriculum, with its implementation of the BEAR Assessment System, to be the only curriculum where the assessment reflected the recent developments.

The IEY course addresses five *progress variables*, which are a major focus of instructional and assessment activities. The progress variables are as follows:

1. **Understanding Scientific Concepts (UC):** Recognizing and applying relevant scientific concepts (e.g. threshold, measurement, properties of matter) to an investigation or problem solution. This variable is the IEY version of the traditional “science content”, although this content is not just “factoids”.
2. **Designing and Conducting Investigations (DCI):** Designing a scientific experiment to answer a question or solve a problem, selecting appropriate laboratory procedures to collect data, accurately recording and logically displaying data (e.g. in graphs and tables), and analyzing and interpreting results of an experiment. This variable is the IEY version of the traditional “science process”.
3. **Evidence and Tradeoffs (ET):** Identifying objective, relevant scientific evidence, and evaluating the advantages and disadvantages of different possible solutions

to a problem based on the evidence available. This variable is unique to IEY, and attempts to capture the cognitive aspects of societal decision-making.

4. Communicating Scientific Information (CSI): Organizing and presenting results, arguments, and conclusions in a way that is free of technical errors and effectively communicates with the chosen audience.
5. Group Interaction (GI): Developing time management skills, the ability to work together with teammates to complete a task (such as a lab experiment) and to share the work of an activity.

Each of these variables is composed of two or more sub-parts known as elements. An example of elements within a variable will be given in the following discussion.

The assessment activities in the IEY curriculum are all designed to assess student performance on one or more of the five variables. Once students have completed an assessment activity, teachers rate student performance into 5 ordered, qualitatively different categories, labeled 0 through 4, using a scoring guide, which is unique to each progress variable, but which was based on the SOLO taxonomy developed by Biggs & Collis (1982). Each scoring guide describes the kind of performance that can be expected from students at the various performance levels. As an example, the scoring guide for the two elements of the ET variable is shown in Figure 1.

Insert Figure 1 about here

The IEY course is divided into four major sections throughout the school year. Each section of the curriculum has a different focus. The first three sections

each focused on a specific scientific issue: (a) Water; (b) Materials Science; (c) Energy. The fourth and final section of the course is a review section which summarizes the material in the previous three sections around an environmental theme. Each of these sections of the course contains a number of assessment tasks, in which at least one and often several of the above progress variables are assessed.

An example² of an assessment task from the IEY curriculum is as follows: Students have been studying issues of water usage and safety. They learn about the cholera epidemic in 19th century London, and how Dr. John Snow went about investigating the cause of the epidemic. They are then asked to write a letter to the editor of a hypothetical London newspaper, pretending that they are Dr. John Snow, and convincing the reader of the cause for cholera, based on the evidence Dr. Snow collected, and recommending what should be done. This activity can be scored on one element of the ET variable, and on both elements of the CSI variable.

To assist teachers in using the scoring guides, a set of exemplars of student work has been assembled. These are drawn from actual student responses. An exemplar for every scoring level of every variable assessed in each activity was chosen. Figure 2 contains exemplars of both a Level 3 and a Level 2 response to the Dr. John Snow letter for the Using Evidence element of the Evidence and Tradeoffs variable.

Insert Figure 2 about here

² A selection of several different kinds of IEY assessment tasks is available in Roberts & Wilson (in press). For a complete listing of all IEY assessment tasks, see Wilson et al (in press).

Assessment of student progress through the course requires that increases in task difficulty, which are a natural part of the curriculum, be disentangled from increases in student proficiency. Therefore we developed a set of 14 unique assessment activities, similar in form to the embedded assessment activities, but which did not require that the student be actively engaged in the curriculum to respond to them, and requiring less time for student response. These activities were divided into five overlapping tests, called link tests, which were given between each of the sections of the curriculum (these were known as link tests I, II, and III) and at the beginning and end of the school year (these were known as the pre-anchor and post-anchor tests). Each of these link tests shared at least one activity with the previous and with the next test, to help establish data links, so that student progress throughout the school year could be evaluated. This bank of assessment activities is also available to teachers who wish to structure more formal end-of-unit tests in addition to ongoing embedded assessment activities. For an overview of the BEAR Assessment System as implemented in IEY curriculum, see Roberts et al (1997).

Development of Progress Maps

Progress maps were developed for four of the five IEY variables. Maps were not developed for the Group Interaction variable, due to the nature of the data that was available to us. These maps are based on the ideas of Rasch measurement (Rasch, 1960; Wright and Masters, 1982), a principal feature of which is that it allows one to simultaneously place both persons and items on the same scale. The technical details of map development are discussed elsewhere (Draney & Peres, in

press). Here, we will focus on the meaning and use of these maps within the context of the assessment system. Because the proficiencies of persons and the difficulties of items are on the same scale, they can be directly compared. This allows us to make statements about the kind of work the student has mastered, the level at which the student is currently working, and the kind of work that is beyond the proficiency of the student. We use these maps to provide a criterion-referenced account of the progress variables.

The map produced for the ET variable is shown in Figure 3. A similar map was produced for each of the other three variables, and for the total IEY Science variable. In this map, the column on the far right contains what we call criterion zones. These criterion zones, indicated by horizontal shaded bands across the map, represent the kind of work that a student located within the bands would typically be doing. If a student falls into criterion zone two, for example, this does not mean that the student will receive a two on every task, but that they would be expected to score 2 on a task of average difficulty. The student may very well receive higher scores on easier tasks and lower scores on harder ones. The criterion zone simply represents average or overall performance, adjusted for task difficulty. The column just to the left of the criterion zone column contains a transformed version of the numerical scale in which item difficulty and person proficiency were originally measured (the original scale, in units called logits, runs from about -3.00 to 3.00). Since teachers and students often find it disturbing for students to receive scores in the negative range, we performed a transformation on the original scores. This does not, however, affect their meaning, in the same way that temperature

measurements can be transformed from Fahrenheit to Celsius, without affecting the meaning of the relationship between two temperature readings. This scale, which we call the SEPUP scale, ranges from 1000 to 2000.

Insert Figure 3 about here

Each of the remaining columns represents a set of activities, or a pretest, link test, or posttest, as indicated by the column label. The activity set and test columns of this map contain all possible raw scores for each activity set or test. For example, in the far left column, representing the combined pre-anchor and pretests, there were 4 ET scorable items. Since each scorable item has a maximum of 4, the maximum sum of all ET scores on the combined pretests is 16. If a student received one score of 2, and three scores of 1 on these tasks, her total score would be 5, as is the case for the student represented on this map. This is called a "raw" score. This student has indicated her performance on the ET variable for the combined pretests by marking 5 in the "Pretests" column (e.g., by circling as in our Figure, or making an X, etc.). This is how the student's exact location on the variable map is determined. Marked in each of these columns is a location on the ET variable for this hypothetical student for each activity set or test.

Once the student's location has been marked on the map, the SEPUP scale and criterion zone information on the right side of the map can be used to interpret the score. For example, the score of 6 on the combined pretests described above corresponds to a SEPUP scale score of approximately 1350, and is near the bottom of the Level 2 criterion zone. We will describe in more detail the many uses of this

information in a moment, but notice the following: Rather than simply saying: "The student received a score of 6 out of 16 on the pretests," a teacher can now say "On the pretest questions, this student typically provided correct information, but left out something important." This is a criterion-referenced interpretation in the original spirit of Glaser (e.g., Glaser, Lesgold & Lajoie, 1987).

The same procedure is then followed at the end of each activity set or test. At the end of Sections A & B of Activities Part 1 (Water), the student would add up the ET scores on all of the activities done in that first section. Just as for the pretests, the student would mark that score in the "Sections A & B of Activities Part 1 (Water)" column, and so on. Since, as part of the IEY curriculum, students are encouraged to keep track of their own progress in various ways, we have tried to make the mapping procedure straightforward enough that the students can do it for themselves.

Correlations between the four SEPUP variables are given in Table 1. Correlations range from .67 to .86. As we might expect, these moderate correlations show that the variables are positively related (i.e., they are all part of what we commonly call "science" in school curricula), but they are also somewhat different, as they should be, given their different definitions.

Insert Table 1 about here

Uses for the progress maps

Uses at the student level

There are a number of uses for the IEY variable maps. As part of the SEPUP curriculum, students often keep notebooks containing their work, scores received on assessment tasks, journal entries, lab notes, and so on. Students can keep the progress maps in this notebook, and mark them as described above. Thus, students can keep track of their own progress on each of the four IEY variables, and overall. Of course, teachers may choose to do this for themselves. They may also choose to have the progress maps printed out by a computer program called *GradeMAP* (Wilson & Draney, in press). Because locations on maps are associated with criterion zones, and not simply numerical scores or percentages, students are more likely to understand what the quality of their current performance is (in terms of the different levels in the scoring guide), and what they need to do to improve it. For example, the student shown in Figure 3 has made steady progress through the first section of the course. She began doing work that was typically either incorrect, or beginning to be correct but with something important missing, and had progressed to the point where her work was, on average, correct and complete. However, for section B of Part 2 of the course, this student's scores have declined significantly, indicating that the student's work is now mostly either incomplete or incorrect. This child's teacher might want to explore with the student, or the student's parents, the reason for this decline. The student may be having problems with this particular content area, or there may be external problems which are affecting the quality of the student's work. The teacher can also discuss with the student the steps

necessary to return to higher levels of performance. We have found that teachers and students can use the maps as part of grading conferences, and the final locations of students, as well as their progress during the year, can be incorporated into teachers' grading schemes.

As another example of diagnostic use, if a teacher notices that an individual student is making good progress on most of the variables, but is having difficulties with, say, Communicating Scientific Information, the teacher can investigate the reasons for the student's particular difficulty. This kind of detailed information is usually not available from simple percentage-correct scores.

Also, variable maps can be used by teachers to communicate with parents during conferences. Rather than simply giving parents the children's scores on tests and percentage of work completed, teachers can use these maps to explain to parents the quality of a student's typical work and the kind of progress the student is making. This could probably best be accomplished using the overall maps.

Uses at the classroom level

Teachers can also use the maps to keep track of the progress of their class overall. Class scores for each test or activity set could be averaged, and the average (rounded to the nearest integer) plotted on a map in the same way as the score of an individual student. Alternatively, each student could be plotted on the same map, to give a picture of the distribution of the entire class, as is shown in Figure 5. These can be done for each variable, and also for the total across all four variables. Such maps can be posted in classrooms so that the students can keep track of the progress

of their class, or used to communicate with school administrators and others interested in the overall performance of science classes.

Insert Figure 4 about here

In Figure 4, a teacher has marked with an X the location of each student in a class of 20 on the first part of the course, and has marked the mean of the class performance with an M (the score numbers have been deleted to make the map easier to read). We can see that students in this class exhibit a range of performances, from students who are producing work which is mostly complete and correct to students who appear to be doing mostly incorrect work. The performance of the class has risen during the year, as shown by the fact that the mean, as well as most of the students, started near the bottom of the map at pretest time, and are in the upper 2 or 3 range at the time of Link Test 2. Such information can also be used diagnostically, both at the group and at the individual level. For example, if the teacher notices a drop in overall group performance after a particular unit, as in Section D of Part 1 in Figure 4, the teacher might want to review the material covered in that unit to make sure that the students understand and feel comfortable with that material.

Evaluation uses of the Maps

The progress maps can also be used as the basis for evaluating educational programs. For example, we examined the effectiveness of the BEAR Assessment System in a field-test of the IEY curriculum in 1994-5. We found differences between growth in different types of science classes, and the criterion-referenced description of the variable developed in the previous sections allows these differences to be substantively interpreted.

The purpose of this field test was three-fold: to gather data about assessment difficulty and student proficiency for all of the tasks in the assessment system, to elicit teacher and student reaction to the assessment system, and to compare the performance of students in classrooms using the curriculum and the assessment system to others using only the curriculum, and to those in more traditional science classes. The field test of the full assessment system took place at 6 centers around the country known as Assessment Development Centers (ADCs). Each ADC consisted of four to six teachers, each of which used the IEY curriculum and the embedded assessment system with at least one middle-school science class. We did not specify anything concerning the selection of students of these teachers, except that they be at an appropriate grade level (7 to 9). In each of these classrooms, data were collected from various assessment activities and from the link tests. In addition, a set of activities was administered in the fall as a pretest, and again in the spring as a post-test. (Note that none of the ADC classes actually completed Part 4, the review portion of the curriculum. Thus, link test 3 was given at approximately the same time as the post-anchor and post-tests.)

In addition, there were seven Professional Development Centers (PDCs) involved in this study, structured similarly to the ADCs, where teachers taught the IEY curriculum and were provided with the same assessment materials as the ADC group, but were not required to use the assessment system. Each Center, both ADC and PDC, was also asked to choose a comparison teacher, who was to be as similar as possible to the other teachers in the center, but who taught the regular middle-school science curriculum. Pretests and posttests were administered to the PDC and comparison groups, but link tests were not. The structure of this data collection is shown in Figure 5.

Insert Figure 5 about here

As is often the case in field studies, not every Center complied completely with these instructions. As is commonly the case with longitudinal data, many students are missing some data, but we did secure relatively complete data from 197 ADC students, 249 PDC students, and 274 comparison students, for a total of 720 students. For all analyses, it was assumed that where data were missing, they were missing at random.

For pretest/posttest group comparisons, the data were analyzed as follows: Only persons with data at both the pretest and the posttest were considered for analysis. Difficulties for pretest items, and abilities for persons, were computed using a modification of the rating scale model on the pretest items (see Draney & Peres (in press) for more information). Item difficulties for the posttest were then anchored to their pretest values, and person abilities computed for the posttest. This

analysis was done for “Total Science” scores (the composite of the four variables). Average abilities and standard deviations for each group at the pretest and posttest, and average gains for each group between pretest and posttest times, are given in Table 2. Gains are shown in SEPUP scale units. Average gains are quite similar for both the PDC and combined comparison groups. The average gain for the ADC group is roughly three times as great as for either of the other two groups.

 Insert Table 2 about here

Table 3 shows the mean and standard deviation of student proficiency in SEPUP scale units for the ADC group at the four testing times (pretests, Link Test 1, Link test 2, and end of year, which includes both Link Test 3 and the posttests). Figure 6 graphically illustrates the progress of the ADC, PDC, and combined comparison groups during the 1994-95 school year, and the relative sizes of the gains from pre to post. The ADC group makes steady progress for the first three time periods. At the time of the posttest, however, there is a significant drop in student proficiency. There are several possible explanations for this. The first is that the posttests were most often administered during the last week or two of the school year, which tends to be a busy time for both teachers and students. It is possible that the administration of some of the posttests was somewhat rushed. Also, as can be seen in the figure, much more time was spent on the first sections of the curriculum than on the last section. In particular, most ADC classes spent nearly half of the school year on Part 1 (Water) of the curriculum, and many spent only a month or so on Part 3 (Energy). Thus, one would expect that students would have felt much

more secure with material from Part 1 than with material from Part 3. Since questions from Part 3 are included in both Link 3 and the post-tests, this may adversely affect the proficiency estimates of students at the time of the posttest. Even with this drop in proficiency, however, it is clear that ADC students made substantial gains during the school year, especially as compared with the PDC and comparison students.

Insert Table 3 and Figure 6 about here

The comparison between the gains made by ADC students and the gains made by PDC students is of particular significance. Both groups used the same curriculum, and the same materials were provided to both groups. Both groups met in teacher meetings throughout the year. The duration of the meetings was the same for both groups. The content was different: the PDC groups critiqued the curriculum materials; the ADC groups participated in local assessment moderation which is discussed in detail elsewhere (Roberts, Sloane, & Wilson, 1996; Roberts, Wilson, & Draney, 1997); briefly, it consisted of the teachers in each ADC meeting about once a month, bringing a set of student papers which all of them had scored, and coming to consensus on the score to be assigned to each paper. Local assessment moderation can serve two purposes: It is a form of scorer calibration, and it can also provide ongoing support for teachers learning to use this new form of assessment.

The primary difference between the ADC and PDC groups was that the ADC teachers paid specific attention to the assessment system: They were required to use the scoring guides to score all of the formal assessment activities, and they

participated in local assessment moderation sessions. This evaluation seems to indicate that the IEY curriculum by itself is not enough to produce the kinds of changes in student performance seen in the ADC group. In addition, a specific focus on assessment is needed to produce this kind of student growth.

Conclusion

These maps illustrate the power of the use of developmental variables and progress maps to enliven the reporting, and enhance the understanding of assessment and evaluation information. The maps that we have shown are only pilot versions for one particular curriculum. We have considerable work left to do in tuning them for teachers, students, and other educational audiences such as parents and school administrators. For example, in order to use the maps when using the pencil-and-paper version of the assessment materials, a teacher must use all of the assessment tasks in a particular task set or link test, or be prepared to provide a “best guess” of student work on missing items. Another limitation is that new tasks developed by the teacher cannot be included in the scores used to produce the maps, since the difficulty of these tasks is not known. In response to these limitations, we are currently pilot-testing software (Wilson & Draney, in press) that will allow the teacher complete freedom in selecting task sets, that will compensate for missing items, and that will print-out individual and class maps.

We have thus far been unable to develop maps for the Group Interaction variable, because it is based on teacher observation of students in groups, rather than on written tasks. We are currently pilot-testing a system, using personal digital

assistants (PDAs), which will allow teachers to record student scores as they observe classroom interaction, and which can also be used to record observations on the other variables. This will make mapping of the GI variable possible, and will also add to the convenience of teacher scoring of tasks.

There are also many technical issues that we have not commented upon here, which will need considerable work in the coming years, including the use of extra information to improve the estimation of student progress, and the linking of IEYP variables to other scales, such as progress variables for other curricula, state and national standards, and NAEP scales. Finally, and most importantly, we are currently undertaking a project to explore more ways that teachers and students can best use this sort of information to enhance learning and instruction.

References

- Biggs, J. B. & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.
- Draney, K. D. & Peres, D. (in press). *Unidimensional and multidimensional modeling of complex science assessment data*. University of California, Berkeley, Bear Research Report SA-99-1.
- Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In, R. Ronning, J. Glover, J. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing*. Hillsdale, New Jersey: Erlbaum.
- Kulm, G. & Malcolm, S. (1991). *Science assessment in the service of reform*. Washington, D. C.: American Association for the Advancement of Science.
- National Science Foundation. (1997, Feb.). *Review of instructional materials for middle school science*. Arlington, VA: Author.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen Paedagogiske Institut. [1980 University of Chicago Press, Chicago].
- Roberts, L., Sloane, K., & Wilson, M. (1996, April). *Local Assessment Moderation in SEPUP*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Roberts, L., & Wilson, M. (in press). *A Sampler of IEY assessment tasks*. University of California, Berkeley, BEAR Report.

- Roberts, L., Wilson, M., & Draney, K. (1997). *The SEPUP Assessment System: An overview*. University of California, Berkeley, BEAR Report.
- Sloane, K., & Wilson, M. (under review). Designing an embedded assessment system: Assessment issues in practice
- Shavelson, R. J. & Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership*, 49,(8), 20-25.
- Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan*, 72, 534-39.
- Wilson, M., & Draney, K. (in press). *GradeMap 1.0* [computer program]. Bear Research Report Series, University of California, Berkeley.
- Wilson, M., Roberts, L., Draney, K., & Sloane, K. (in press). *SEPUP Assessment Resources Handbook*. SEPUP, Lawrence Hall of Science, University of California, Berkeley.
- Wilson, M. & Sloane, K. (under review). *From principles to practice: An embedded assessment system*.
- Wolf, D., Bixby, J., Glenn, John, III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education*, 17, 31-74.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Table 1
Correlation between the SEPUP variables

	DCI	UC	CM
ET	.76	.83	.86
DCI		.85	.69
UC			.76

Table 2
Means and standard deviations for three groups at all testing times

	ADC	PDC	Comparison
Pretest	1386 (50.0)	1373 (36.9)	1368 (60.0)
Posttest	1431 (56.3)	1386 (36.9)	1381 (61.9)
Gain	45	13	13

Note: Parenthetical values are standard deviations.

Table 3
Means and standard deviations for the ADC group at 4 testing times

Time	Mean	SD
Pretest	1386	50.0
Link 1	1421	65.6
Link 2	1476	68.8
Link 3 & posttest	1431	56.3

Figure Captions

Figure 1. The scoring guide for the Evidence and Tradeoffs variable.

Figure 2. Exemplars for Levels 2 & 3 of the Evidence and Tradeoffs variable at the Dr. John Snow letter.

Figure 3. Student map for the Evidence and Tradeoffs variable.

Figure 4. Class map for the Evidence and Tradeoffs variable.

Figure 5. Structure of the 1994-95 SEPUP data collection.

Figure 6. Gains for the ADC, PDC, and comparison groups on the combined SEPUP variable during the 1994-95 school year.

Score	<p align="center">Using Evidence:</p> <p>Response uses objective reason(s) based on relevant evidence to support choice.</p>	<p align="center">Using Evidence to Make Tradeoffs:</p> <p>Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice.</p>
4	Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence.	Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence & explaining how it influences choice.
3	Response provides major objective reasons AND supports each with relevant & accurate evidence.	Response discusses <u>at least two</u> perspectives of issue AND provides objective reasons, supported by relevant & accurate evidence, for each perspective.
2	Response provides <u>some</u> objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete.	Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete & accurate perspective has been provided.
1	Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity.	Response states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	No response; illegible response; response offers no reasons AND no evidence to support choice made.	No response; illegible response; response lacks reasons AND offers no evidence to support decision made.
X	Student had no opportunity to respond.	

Maps for student progress

Level 3	Level 2																									
<p>Dear London Health Department:</p> <p>I have gone to a lot of trouble to find the cause of the new outbreak of cholera here in London. My belief is that bits of human waste in the drinking water is causing the spread of cholera. I have been researching about the water and my friend John Joseph Whiting has gone door to door to find the number of fatalities of people drinking water from each company. I know that there are two main water companies that supply the area of the latest outbreak, S & V and Lambeth. They both supply for all kinds of people. The Lambeth water company used to get its water from the same sewage infested area as S & V. But, a few years ago they switched their intake upstream and there was a decrease in the number of fatalities. look at the chart below to see the percentages of fatalities for households served by each water company.</p> <table style="width: 100%; border-collapse: collapse; margin: 10px 0;"> <thead> <tr> <th style="text-align: left;">Water Companies Households</th> <th style="text-align: right;"># of Houses</th> <th style="text-align: right;"># of Fatalities</th> <th style="text-align: right;">% of Fatalities</th> <th style="text-align: right;">% per</th> </tr> </thead> <tbody> <tr> <td>S & V 3.15%</td> <td style="text-align: right;">40,046</td> <td style="text-align: right;">1,263</td> <td style="text-align: right;">45%</td> <td></td> </tr> <tr> <td>Lambeth</td> <td style="text-align: right;">26,107</td> <td style="text-align: right;">98</td> <td style="text-align: right;">4%</td> <td style="text-align: right;">.37%</td> </tr> <tr> <td>All Others .55%</td> <td style="text-align: right;">256,423</td> <td style="text-align: right;">1,422</td> <td style="text-align: right;">51%</td> <td></td> </tr> <tr> <td>Total</td> <td style="text-align: right;">322,576</td> <td style="text-align: right;">2,783</td> <td style="text-align: right;">100%</td> <td></td> </tr> </tbody> </table> <p>As you can see, the number of fatalities per household for S & V were much higher than for Lambeth. That helps me prove my idea that it is the water that is spreading the cholera.</p> <p>You should also inform the people that use S & V which I would be glad to help you do, to not use their water. They should go somewhere that Lambeth provides to get their water. Maybe you could place an inner city pump with non infected water if S & V does not cooperate. I hope you take my ideas into consideration. Thank you for your time.</p> <p>Sincerely,</p> <p>John Snow, M.D.</p>	Water Companies Households	# of Houses	# of Fatalities	% of Fatalities	% per	S & V 3.15%	40,046	1,263	45%		Lambeth	26,107	98	4%	.37%	All Others .55%	256,423	1,422	51%		Total	322,576	2,783	100%		<p>Dear Dr. Snow,</p> <p style="text-align: center;">I am a victim of the Cholera epidemic. I think cholera is spread through the water we drink. S & V water company supplied 40,096 homes with water where 1,263 people died. Lambeth water company supplied 26,107 homes with water where 98 people died. Other water companies supplied 256,423 homes with water where 1,422 people died. Lambeth is the safest water company to get water from. The Health Department should make everybody drink Lambeth's water, but that is not perfectly safe either. 2,783 people died out of 32,576. I think every one in London should go to a different town to stay away from cholera.</p>
Water Companies Households	# of Houses	# of Fatalities	% of Fatalities	% per																						
S & V 3.15%	40,046	1,263	45%																							
Lambeth	26,107	98	4%	.37%																						
All Others .55%	256,423	1,422	51%																							
Total	322,576	2,783	100%																							

Maps for student progress

Pre-Tests	Part 1: Water				Part 2: Materials Science		Part 3: Energy			Post-Tests	SEPU P Scale Score	Developmental Levels
	A & B 1-12	C 13-20	D 21-28	Link 1	A 29-38	B 39-46	Link 2	47-58	Link 3			
											2000	
											1950	
						16	12				1900	Level 4 <i>Goes beyond Level 3 in significant way</i>
			8					16			1850	
		8			12	15	11			20	1800	
16	12			19		14		15			1750	
		7	7	18	11	13	10	14			1700	
15	11			17			9			19	1700	
						12				11		
											1650	Level 3 <i>Correct & complete</i>
14	10	6	6	16	10	11	8	13		18	1650	
				15							1600	
					9	10	7	12		10	1600	
	9			14		9	6				1550	
		5		13	8	8			11		1550	
	8		5	12			5			9	1500	
		4		11	7	7	4	10			1500	
	7		4	10		6		9		8	1450	
		3		9	6	5	3	8		7	1450	
	6			8	5	4				6	1400	
			3	7		3	2	7		6	1400	
	5			6	4			6		9	1350	Level 2 <i>Correct but important part missing</i>
	4	2		5	3	2	1	5	5	8	1350	
			2	4				4	4	7	1300	
	3	1		3	2	1	0	3		6	1300	
			1							5	1250	
	2	0		2	1	0		2		4	1250	
										3	1200	
	1		0	1	0			1		2	1200	
									1		1150	
	0			0				0		0	1100	Level 1 <i>On task but incorrect</i>
											1050	
												Level 0 <i>Off task or Missing</i>

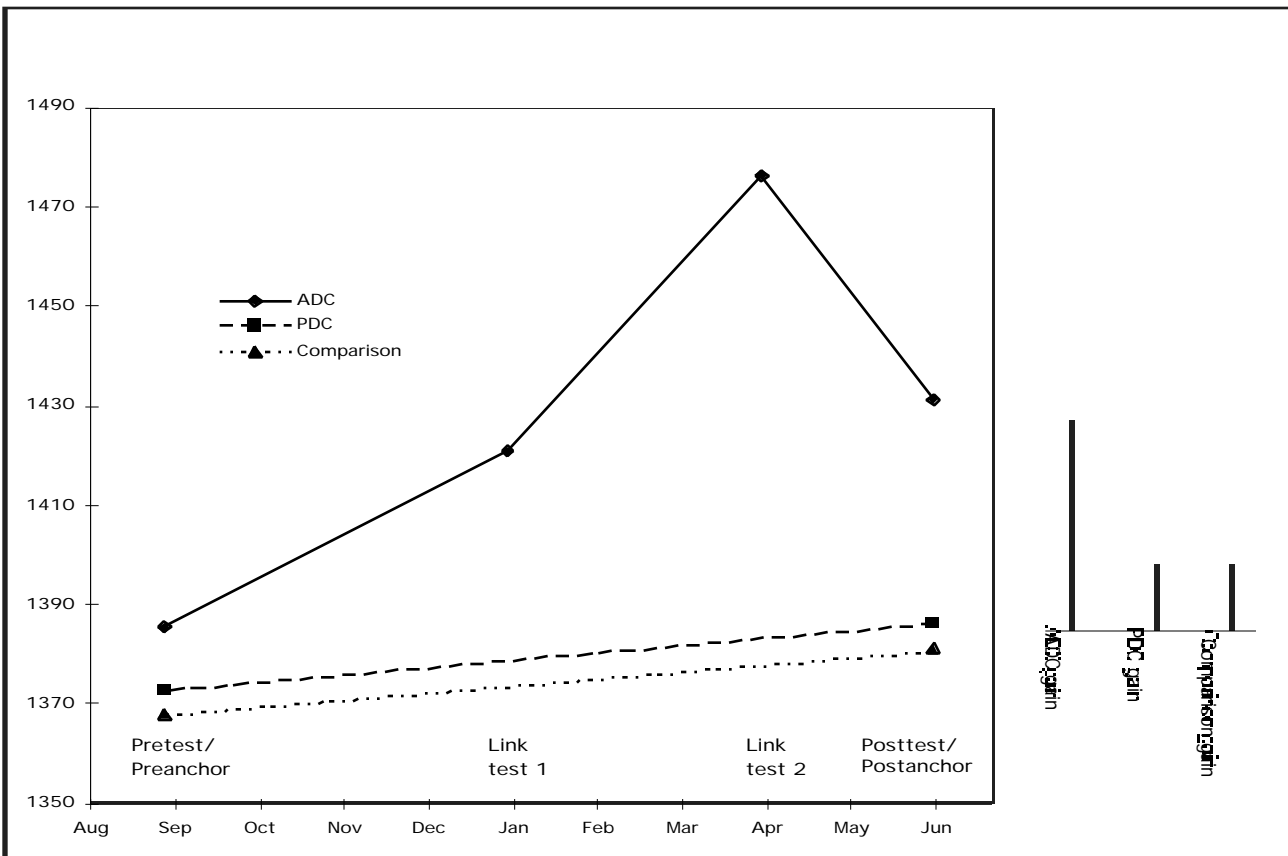
Maps for student progress

Group	September	Part 1		Part 2		Part 3		June
		Water		Materials Science		Energy		
ADC	Pre-anchor							Post-anchor
	Pre-test	Activities	Link 1	Activities	Link 2	Activities	Link 3	Post-test
PDC	Pre-test							Post-test
Comparison	Pre-test							Post-test

Maps for student progress

Pre-Tests	Part 1: Water				Part 2: Materials Science			Part 3: Energy		Post-Tests	SEPUP Scale Score	Developmental Levels
	A & B 1-12	C 13-20	D 21-28	Link 1	A 29-38	B 39-46	Link 2	47-58	Link 3			
											2000	
											1950	
											1900	
											1850	Level 4 <i>Goes beyond Level 3 in significant way</i>
											1800	
											1750	
	X										1700	
											1650	
											1600	Level 3 <i>Correct & complete</i>
		XX	X								1550	
	XX	XX									1500	
	XX	XX									1450	
	XX	XX									1400	
											1350	Level 2 <i>Correct but important part missing</i>
											1300	
											1250	
											1200	
											1150	
											1100	Level 1 <i>On task but incorrect</i>
											1050	
												Level 0 <i>Off task or missing</i>

Maps for student progress



progress

Maps for student