Evolution of Life Science Assessments for Middle School

Barbara Nagle and Marcelle A. Siegel
SEPUP
Lawrence Hall of Science
University of California, Berkeley
Berkeley, CA 94720-5200

Ann Barter
Center for Research, Evaluation and Assessment
Lawrence Hall of Science
University of California, Berkeley
Berkeley, CA 94720-5200

Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (NARST), Vancouver, BC, April 1-3, 2004

Evolution of Life Science Assessments for Middle School

This paper explores the challenges of creating effective extended items to assess middle school life science students' understanding of concepts related to the National Science Education Standards for Inquiry, Life Science, and Science in Personal and Societal Perspectives. We describe our process for developing and revising items in order to enhance opportunities for students to demonstrate their understanding. Our findings suggest that items were improved through a cycle of expert review, revision, and classroom pilot testing. We provide examples of revised items that result in better opportunities for students to demonstrate what they know and are able to do. We also provide examples of typical students' responses to these items.

Barbara Nagle, Marcelle A. Siegel, and Ann Barter
Lawrence Hall of Science, UC Berkeley
Berkeley, CA 94720

The National Science Education Standards (NSES) call for changing emphases in assessment, including a shift from assessing discrete knowledge, what is easily measured, and what students *don't* know to assessing rich, well structured knowledge, what is most highly valued, and what students *do* understand (NRC, 1996). Curriculum developers are now including in their development process this new approach to assessment of student understanding consistent with these changing emphases recommended by assessment experts (e.g., NRC, 2001). New assessment approaches are necessary in the current climate of accountability and high-stakes testing as traditional assessment procedures may fail to capture the full range of student understanding. This shortcoming may have serious policy implications when decisions are made about science curriculum based on incomplete understanding of the impact of the curriculum on student learning.

This paper explores the challenges of creating extended response items that provide appropriate prompts for students to use their understanding and critical thinking or reasoning skills to respond to a complex problem or question intended to elicit rich, well structured knowledge. These items were developed as part of a pre-/post -test for summative assessment of student learning in an innovative yearlong middle school life science course. This course includes an embedded authentic assessment system based on five key assessment variables and associated scoring guides (See appendix for examples of scoring guides developed for these variables.). Embedded assessments differ from traditional assessment of student learning in that the assessment tasks are embedded within daily classroom work. Authentic assessments are based on tasks that are similar to those carried out by scientists and carried out by students as part of instruction. This assessment system provides valid and reliable information about students' abilities to perform complex tasks and provide extended responses to complex questions (Wilson and Sloane, 2000). The summative assessment was designed to measure a range of content and assessment targets (e.g., Stiggins (2001) so it included multiple choice, short answer, and extended response items. The targets assessed in this study include basic knowledge and understanding of life science concepts, reasoning proficiency, and scientific performance skills. In this paper, we will focus on the development of two extended items that were scored with scoring guides (rubrics).

Objectives

This study is part of a larger effort to evaluate the effectiveness in promoting student learning of an innovative NSF-funded seventh grade life science course called *Science and Life Issues*. This larger evaluation addresses the question: *What are students learning in relation to the instructional materials and the related National Science Education Standards?* The specific objective of this study was to develop and characterize developmentally appropriate assessment items of high technical quality for summative evaluation of the course. The extended response items are intended to elicit extended responses that require students to apply scientific processes and concepts and, in some cases, evaluate evidence related to a scientific question or issue.

The National Science Education Standards (NRC, 1996) set forth several criteria for assessment data of high technical quality. The criteria that apply at the task (item) level include the following recommendations:

- the content and form of the task must be congruent with what is supposed to be measured,
- the assessment tasks should be authentic, and
- tasks must be developmentally appropriate, set in contexts that are familiar to the students, must not require inappropriate levels of vocabulary or reading skills, and must be free of bias (NRC, 1996).

Based on these criteria, the following specific questions were posed about the items:

1) Did students answer the intended question? If not, what kinds of problems did they have?
2) Was there a range of responses over the levels of the five-level point scoring guides used to interpret the responses? In other words, did the question provide an opportunity for students to demonstrate their level of understanding and critical thinking?
3) How could the questions be revised in order to provide an opportunity for students to demonstrate what they have learned?

Theoretical underpinnings

The assessment triangle of cognition, observation, and interpretation (NRC, 2001) and an embedded authentic assessment system developed in collaboration with the Berkeley Evaluation and Assessment Research (BEAR) group (Wilson and Sloane, 2000) provide the theoretical framework for the development of tasks. The basic assessment system includes three components, each related to the assessment triangle:

- Cognition: five **variables**, grounded in a developmental perspective of student learning, that define the knowledge and skills in which students are expected to make progress during the year,
- Observation: **assessment tasks** used to collect multiple observations of student performance
- Interpretation: **scoring guides** used to interpret student responses and evaluate performance on the tasks.

For the two items under consideration in the present study, three of the five variables were included; one to assess understanding of scientific concepts and two to assess science process skills, as shown below. These variables align with assessment targets described by other researchers. For example, the variables align with three of the five targets described by Stiggins (2001); knowledge and understanding, (the UC variable), performance skills (the DCI Variable), and reasoning proficiency (the ET variable). The tasks also align with specific learning goals in the instructional materials related to the NSES for life science content, inquiry, and science in personal and societal perspectives. Extended response questions were designed to assess three of the five SEPUP variables as follows:

Scientific Process:

DESIGNING AND CONDUCTING INVESTIGATIONS (DCI)—Designing a scientific experiment, performing laboratory procedures to collect data, recording and organizing data, and analyzing and interpreting the 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.

Scientific 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.

The extended response items are designed to allow students to respond to complex situations, and, in some cases, can be assessed on more than one variable. Achievement targets were selected from key concepts of the course that were judged by the developers to have been addressed thoroughly through several course activities. The two items presented in this paper were developed to assess content emphasized in the course and related to middle level NSES for which no released items are available from either the Third International Mathematics and Science Study (TIMSS) or the National Assessment of Educational Progress (NAEP).

Each variable scoring guide provides criteria for assessing student performance at each level. Although each scoring guide offers specific criteria, all are similar in that a 0 indicates an off-topic or missing response, a 1 indicates an attempt at a response, a 2 is partially correct, 3 is complete and correct, and a 4 goes above and beyond the expected complete response for answering the question, for example a student may after answering the question compare his/her thinking with another similar situation or context.

Design and Procedure

Items were designed by the curriculum developers to correlate to the *Science and Life Issues* yearlong course and to the National Science Education Standards for life science in grades 5–8. They were developed through a rigorous process of planning, item development, pilot testing, and preparation for scoring. The items were initially developed to assess key concepts set forth in the curriculum and correlated to the NSES. They were then reviewed by the developers, collaborating researchers, and teacher associates for age appropriateness, clarity, and content. They were then pilot tested, revised based on analysis of students' responses and additional review by science education experts, and piloted again in classrooms using the curriculum.

The specific items discussed in this paper relate to the NSES for Inquiry, Life Science Content, and Science in Personal and Societal Perspectives. In relation to the Abilities Necessary to Do Scientific Inquiry, the course provided multiple opportunities in a variety of contexts for students to design investigations, including a focus on reproducibility, sample size, and control of variables. The specific life science content standards relevant to the items discussed in this paper are those for Reproduction and Heredity, Structure and Function in Living Systems, and Diversity and Adaptation of Organisms. The specific content related to each item is discussed with the results for the two items discussed in detail in the Findings section of this paper. In relation to the Science in Personal and Societal Perspectives standard, students had multiple opportunities to apply scientific evidence to personal and societal decisions related to course topics such as infectious diseases, genetic conditions, and environmental concerns.

The test was administered shortly before the end of the school year in seventh grade classrooms using the course. In 2000, participating classrooms were from eight national field test centers using the course. In 2001, pilot testing of the test was conducted with 41 students in the classrooms of two teachers implementing the course in two different districts (one suburban, one mixed suburban and rural). In 2002, an expanded pilot test was conducted with 600 students in four districts (one urban, one small city, one suburban, and one mixed suburban/rural). Participating classrooms were selected based on their implementation of a significant portion of the course and their use of the embedded authentic assessment system.

Scorers were trained through the process of assessment *moderation* (e.g., Roberts, Sloane, & Wilson, 1996), in which three or more raters scored a sample of papers independently and then met to discuss the scores and use the scoring guide criteria as the basis for supporting or adjusting the scores until they achieved consensus. During moderation, the raters also discussed the nature of student responses, the degree to which the items were eliciting appropriate responses from students, and possible revisions to the items that would make them easier for students to interpret. The raters also discussed the scoring criteria to be sure that they were clear on the expectations for each item. This process also led to the development of a detailed rubric for each question. Each item was then scored by a single scorer. The development of the detailed rubrics and the methods for training scorers are discussed in detail elsewhere (Siegel, Nagle, & Barter, 2004). Qualitative review of student work and development of codes for student responses was conducted by a course developer.

Additional details of the design are described as appropriate in the following section.

Findings

Item 1: Flower item

The original version of the item shown below was intended to assess both inquiry and life science content knowledge. It addresses the NSES middle level Inquiry content standard for Abilities Necessary to Do Scientific Inquiry by asking students to design an experiment to investigate a hypothesis or prediction. It also addresses the middle level Life Science content standard on Reproduction and Heredity that states: "The characteristics of an organism can be described in terms of a combination of traits. *Some traits are inherited and others result from interactions with the environment*." The item was intended to elicit either controlled investigations of the effects of the two environments on seeds from the two different plants or

breeding experiments. In the course, students investigated the influence of both heredity and the environment on traits of humans, animals, and plants.

Original version (not administered to students)

Two scientists discover a new kind of flower in a remote area. In a damp field on the side of a stream, they find yellow flowers. In a rocky area on the side of a nearby mountain, they find a flower that appears identical to the first flower, except that it is orange. One of the scientists thinks the color difference is due to the different locations where the plants are growing. The other thinks it is a genetic difference.

Design an experiment to gather evidence to help decide which scientist is correct.

This original version was revised, as shown below, based on developer review. The developers thought the item, with two contrasting hypotheses, would be too complex for seventh grade students. However, this change (shown below) focused the item on inquiry only, and removed the focus on reproduction and heredity. The following version was administered to students during a field test of the course. The changed element of the question is underlined.

First revision (2000)

Two scientists discover a new kind of flower in a remote area. In a damp field on the side of a stream, they find yellow flowers. In a rocky area on the side of a nearby mountain, they find a flower that appears identical to the first flower, except that it is orange. <u>One of the scientists thinks the color difference is due to the different locations where the plants are growing.</u>

Design an experiment to gather evidence to help decide <u>if the</u> scientist is correct.

The student papers were not scored, but a sample of about 100 was reviewed to determine whether students were able to make reasonable attempts at answering the question. Over half of the responses to this item included a suggestion of an experiment or study. Most of the students who answered the question wrote one-sentence descriptions of experiments that involved either switching locations or controlling conditions such as moisture or soil. The biggest problem was that many students did not answer the intended question by suggesting an experiment, but rather offered hypotheses or opinions that agreed with (or occasionally disagreed with) the hypothesis presented in the question. The overall results suggested that the item was promising, but needed to be revised to prompt more complete answers and encourage students to propose experiments, rather than hypotheses or opinions.

In order to provide additional cues that the question required an experiment and to encourage students to think systematically about the problem and provide more evidence about their thinking, the requirement to include a data table was added to the second revision, shown below. In addition, a second part was added to the question to determine whether students could identify alternative hypotheses for the observations in the prompt. This addition was suggested by the fact that a very small number of students in the 2000 study suggested breeding experiments or the hypothesis that the different flower color was inherited or intrinsic to the plant. The use of additional questions in the same prompt reduces the need to set up additional scenarios, which often take students significant time to

read and understand. With the additional question, the item provides an opportunity to probe students' understanding of the role of heredity and the environment in determining an organism's traits.

Second revision (2001)

Two scientists discover a new kind of flower in a remote area. In a damp field on the side of a stream, they find yellow flowers. In a rocky area on the side of a nearby mountain, they find a flower that appears identical to the first flower, except that it is orange. One of the scientists thinks the color difference is due to the different locations where the plants are growing.

a. Describe an experiment to gather evidence to help decide if the scientist is correct. Include a data table for recording your results.

b. Suggest an alternative hypothesis that might explain the different flower colors. Be sure to explain your hypothesis.

These items were scored: Part a was scored with the Recording Design or Procedure element of the Designing and Conducting Investigations scoring guide (see p. 17 of the Scoring Guide Appendix). Part b was scored with the scoring guide for Understanding Concepts (see p. 18 of the Scoring Guide Appendix).

Review of the responses to this question suggested that the mention of a data table and results in the prompt led to an increase in the number of students proposing experiments even though many students still neglected to include a data table.

The results of the scoring are shown in the 2001 rows of Table 1 on the following page. The results show that there was a distribution of the scores over the 4-point scoring levels for both parts of the item. In some cases, an intermediate level (such as 1.5) was added when it was possible to distinguish responses that had all of the requirements for one scoring level, but only some of the requirements for the next scoring level. For the purpose of this qualitative study, results are compared only in light of changes to the item itself and not to the mean score of the answers. Responses are provided as percentages to help observe general trends. Statistical comparisons of mean scores for items for different year administrations are not appropriate (given the item has changed and student and teacher populations changed each year).

Table 1: Flower item scores

Date of administration of the item	N	Score Levels									Mean score/total possible
		0	0.5	1	1.5	2	2.5	3	3.5	4	
2001/a (expt & table)	41	5	0	8	9	12	3	2	0	2	1.63/4
2002a (expt only)	84	18	11	24	0	9	19	3	0	0	0.99/4
2002b (table only)	84	15	10	3	17	36	0	3	0	0	1.36/4
2001/b (hypothesis)	41	8	0	7	6	7	0	12	0	1	1.77/4
2002/c hypoth 1	84	9		75							0.89/1
2002/d hypoth 2	84	21		63							0.75/1

Note that in 2001, part a included both the experiment (expt) and table, while in 2002 these items were separated into a, the experiment only, and b, the table only. Also, in 2001 part b asked for a hypothesis and explanation, while in 2002 part c asked for a hypothesis and part d asked for as many more additional hypotheses as students could think of.

Based on scoring and moderation of a sample of the student responses, the item was further revised, as shown below. Part a of the question was divided into two parts: a for the experiment and b for a data table. In addition, the hypotheses students generated in part b of the question usually focused on other environmental differences. In order to probe further to see if students might come up with an alternative hypothesis involving heredity, this part of the question was also split into two parts, one requesting a hypothesis and the second requesting *as many additional* hypotheses as students could think of. Note also the underlined changes, made to focus students in their design on the variable of water and to more dramatically distinguish the flower colors.

Third revision (Spring, 2002)

A scientist discovers a new kind of yellow flower in a damp field on the side of a stream. In a dry rocky area on the side of a nearby mountain, the scientist finds a flower that appears identical to the first flower, except that it is red. The scientist thinks the color difference is due to different amounts of water in the soil where the plants are growing.

a. Describe an experiment to gather evidence to help support or disprove the scientist's idea.

b. Include a data table for recording your results.

c. Suggest an alternative hypothesis that might explain the different flower colors. Be sure to explain your hypothesis.

d. List as many more reasonable hypotheses as you can think of.

Part a of this question was scored with the Recording Design or Procedure element of the Designing and Conducting Investigations scoring guide, while b was scored with the Organizing Data element of the same scoring guide. Parts c and d were scored either 0 or 1, based on whether students supplied at least one reasonable hypothesis for each of these two parts of the question. This change in scoring reflected our decision to focus on whether students could generate reasonable hypotheses, rather than whether they provided full explanations of their hypotheses. The results shown in the 2002 rows of Table 1 illustrate the range of student responses.

Due to the splitting of the item, the scoring criteria for the experiment and data table became more rigorous, so the scores from 2001 to 2002 cannot be compared to determine the item effectiveness. In addition, the hypothesis question was changed from one 4-point question to two 1-point questions, reflecting a correct/incorrect response. Thus it was necessary to develop other approaches to coding responses in order to compare the items from 2001 and 2002.

Answers were reviewed to determine how many student responses included an experimental approach and how many included a data table. The experimental responses were further coded based on whether they involved switching the plants' environments (S), controlling the amount of water plants received (W), or other (O) types of experiments or investigations such as testing the soil, measuring the water, breeding the plants, or testing the plants' DNA. The data tables were coded based on whether they were attempts with little logic (A), showed some logical thinking (S), or exhibited sound logic (L). The results are shown in Table 2: Flower item: results for experiment and data table.

Table 2: Flower item: results for experiment and data table

		Part a: Experiment				Part b: Data Table	
Date of item administration	N	Percent with an experiment	Percent suggesting switching (S)	Percent suggesting controlling water (W)	Percent suggesting other (O) approaches	Percent providing a data table	Percent providing somewhat logical /logical data table
2001	41	83	51	0	32	34	17/10
2002	84	77	33	27	17	85	52/23

In 2001, a total of 83% suggested some type of experiment or investigation, with 51% suggesting switching the flowers (or their seeds) into different environments, and 32% suggesting other approaches such as studying the soil or water in the two locations, studying the plants themselves, or testing the plants' DNA. This latter response, given by 2 students (5%), indicated that these students were aware that heredity provides an alternative hypothesis for the flower results suggested in the item. However, only 34% included a data table in their responses and only about 1/3 of these data tables were complete and logical.

When the item was revised for 2002, 77% of the students included an experimental approach in part a. This is slightly less than the 83% who did so the previous year.

However, the addition of part b suggesting a data table resulted, as expected, in a large increase in the number of students who included one. Now 85% of students provided a data table. The total percent of students with somewhat logical or logical data tables increased from 27% to 75%.

A sample level 3 student response for each part of the question follows.

> *Response to part a: "Plant 50 flowers in a dry area, and plant 50 flowers in a wet area. Put 25 each of red and yellow in each bed. Observe color of blooms."*
>
> *Response to part b:*

	color of flower seed from	*color of seeds flower*
lot of water	*red*	
	yellow	
little water	*red*	
	yellow	

The 2001 and 2002 responses were also reviewed to determine how many students included alternative hypotheses based on the environment and how many were based on heredity or other characteristics intrinsic to the plant itself. The percentage of students with each type of response is shown in Table 3: Flower item: results for alternative hypotheses.

Table 3: Flower item: results for alternative hypotheses

Date of item administration	N	Percent environmental hypotheses	Percent genetic or other "intrinsic to the plant" hypotheses
2001	41	49	40
2002	84	56*/74**	32*/56**

* include this type of hypothesis in response to question c
** indicate this type of hypothesis at least once (on either c or d): note that total percentage of nature + nurture exceeds 100 since some students provide multiple hypotheses

In 2001, 40% of students suggested non-environmental hypotheses. This includes students who explicitly mentioned mutations or dominant vs. recessive traits, referred to "pollen mixing" or mating, or suggested that flower color was "in the seed." The

remaining non-environmental hypotheses suggested that they were different plants or different color variations of the same plant, "like roses."

The 2002 revision was effective in eliciting more hypotheses related to heredity or other characteristics intrinsic to the plant or species. The total percentage of students with non-environmental hypotheses for part c was 32%, similar to that for the previous version of the item. However, when students who added genetic or evolutionary hypotheses in part d are included, the percentage of students with these hypotheses rises to 56%. The variety of these responses also increased, to include mentions of inheritance, different genes, mutations, different breeds, adaptations, natural selection, evolution, and different species. The number and variety of environmental hypotheses also increased. Sample environmental hypotheses provided by students in response to parts c and d of the 2002 flower item include:

> *"The flowers could be different colors depending on the weather."*
>
> *"There could be certain colors of dye in the ground which changes the color of the flowers."*
>
> *"An alternative hypothesis is that the yellow flower in the field gets more sunlight."*
>
> *"Different amount of sunlight, different soil, more erosion, different flowers that adapt differently."*
>
> *"Maybe the flowers are that color because of genes that are passed on. The red flowered plants would be in that area because that is where its "parents" were."*
>
> *"It's just another flower of the same type but different color."*
>
> *"It could just be a variation, like animals are the same breed but have some difference."*
>
> *"Age, mutation, gender."*
>
> *"A mutation could have happened because they both looked the same but different color."*

The analysis of students' responses to the flower item indicates that, as expected, students provide more complete responses when the question is broken down. It also shows that the majority of students are able to answer the intended questions to design an experiment and data table and to suggest alternative environmental and genetic or evolutionary hypotheses. However, most students' experimental designs and data tables are judged to be only partially correct. Two additional items—one multiple choice and one 2-point free response item—have been designed to determine whether students' understanding of heredity and the environment in determining an organisms' traits is stable. These items are currently being analyzed.

Based on the review of students' responses to the Spring 2002 version of the item, it was slightly revised for a 2002–2003 pretest/posttest, as follows. The results from this administration of the test are currently being analyzed.

Fourth revision (administered in June, 2003)

> A scientist discovers a new kind of yellow flower in a damp field on the side of a stream. In a dry rocky area on the side of a nearby mountain, the scientist finds a flower that appears identical to the first flower, except that it is red. The scientist thinks the color difference is due to **different amounts of water in the soil** where the plants are growing.
>
> a. Describe an **experiment** to gather evidence to help support or disprove the scientist's idea.
>
> b. Include a data table for recording your results.
>
> c. Suggest an alternative hypothesis that might explain the different flower colors. Be sure to explain your hypothesis.
>
> d. List as many more reasonable hypotheses as you can think of.

Item 2: Antibiotic resistance

The following item was developed to determine whether students understood the concept of antibiotic resistance and whether they could discuss it in the context of a decision about whether to stop taking antibiotics due to a mild side effect. The curriculum included a simulation and reading on the emergence of antibiotic resistance with a focus on variation in a population of bacteria. This was part of a unit on the cellular nature of life and infectious diseases, but before the unit on evolution and natural selection. The question was scored with both the Understanding Concepts and Evidence and Trade-offs scoring guides on pp. 18 and 19 of the Scoring Guide Appendix. This question also went through several revisions. Review of student responses suggests that the revisions are resulting in more complete answers, although analysis is not yet complete.

> **2000 version**
> Rita gets ear infections quite frequently. When she does, her doctor has her take an antibiotic for ten days. Three days ago, Rita got an ear infection. This time, the antibiotic worked really quickly, and Rita felt completely better after only two days. Her parents want to know if she can stop taking the antibiotics.
>
> a. Should Rita stop taking the antibiotics? (Answer "yes" or "no.")
>
> b. Explain your answer to Question 25(a).

> **2001 version**
> Rita began taking a ten-day treatment of antibiotics three days ago. The antibiotics worked quickly, and Rita feels completely better after only three days. Antibiotics upset Rita's stomach, so she wants to stop taking them.
>
> Should Rita stop taking the antibiotics or finish the treatment? Explain the advantages and disadvantages of stopping and of continuing the antibiotics. Be sure to include your final recommendation and reasons.

2002 version
Rita began taking a ten-day treatment of antibiotics three days ago. The antibiotics worked quickly, and Rita feels completely better after only three days. Antibiotics upset Rita's stomach, so she wants to stop taking them.

Should Rita stop taking the antibiotics or finish the treatment? Explain the advantages and disadvantages of stopping and of continuing the antibiotics. *Be sure to include your final recommendation, any trade-offs involved, and your reasons for your decision.*

This item can be scored independently for the two different scoring guides. Some sample answers follow.

Level 2 for UC, level 2.5 for ET:

"If Rita stops taking the antibiotics, then she won't get better. If Rita continues to take the antibiotics then the medicine will upset her stomach. Rita should keep on taking the antibiotics but should try other medicines to stop upsetting her stomach.

Level 3 for UC, Level 2 for ET

"No, she should continue taking them otherwise the bacteria left would become resistant and harder to cure after they come back and she takes a turn for the worse."

Level 3 for UC, Level 3 for ET

"Rita should not stop taking the medication!! If she stops too early, she could then have an antibiotic resistant strain of bacteria in her. If she keeps taking the antibiotic it will most likely kill off the bacteria. The only trade-off is that her stomach will be upset for a few days, but what's that to bacteria? My final decision is that she shouldn't stop taking the medication."

Conclusions and Discussion

This study has resulted in the development of a number of items on concepts that relate to the seventh grade curriculum and NSES content standards for life science, inquiry, and science in personal and societal perspectives. As an example, the 2002 version of the flower item, presented in depth in this paper, appeared to be effective in eliciting students' ability to design a controlled experiment, create an appropriate data table, and generate alternative hypotheses related to the role of heredity and the environment in determining an organism's traits. The antibiotic item was designed to elicit students' understanding of antibiotic resistance and ability to analyze the trade-offs of a decision about whether to discontinue the antibiotics due to a mild side effect. Not all items developed in a positive direction: some have been discarded or changed to a 0-1 or 0-1-2 scoring system.

Although the findings are preliminary, they illustrate principles that may be helpful to other educators and researchers interested in obtaining high quality data about student learning at the middle school level. These relate to development of the item and to scoring.

When developing items for seventh graders, breaking the item into small pieces that call for each type of desired response increases the likelihood of students attempting each part of the response. In the flower item shown, separating the graph from the experimental design encouraged more students to complete all required components of the item. This suggests that students either become impatient with reading longer items or that they simply lose track of all of the demands of the item. Sound items should specify the knowledge and kinds of reasoning expected and point the direction to an appropriate response without giving away the answer (e.g., Stiggins (2001). Sometimes, these elements can be incorporated into separate parts of the question. For example, the prompt to include a data table in the flower item does not give away the answer in any way, but points students to designing an experiment and specifies the expected approach. (It seems to get lost when not broken out from the rest of the question.) Similarly, the addition of prompts to include advantages and disadvantages and trade-offs of a final decision to the antibiotic item improves the item.

Despite a focus in the course materials on writing thorough answers and assessing them with the scoring guides, most students wrote only one to three sentences in response to each item. This is another reason why a given item is likely to elicit more response if broken into smaller pieces. It also suggests that some items should be designed to only require a sentence or two. As a result of this finding, we have developed many more short answer items scored on a 0-1 or 0-1-2 basis in order to explore student understanding and learning. Some of these short answer items are independent (that is, not related to any extended response item), while others are parts of extended response items like the flower item, which includes a mixture of 0–4-point sections and 0–1-point questions.

Another approach suggested by this research is to add a question pushing students to go further in their response, as in part d of the flower item that asks students to think of *as many other hypotheses as possible*. Perhaps the experimental design in part a might be followed by another question: *Review your answer to part a. Think about the characteristics of a good experimental design. How could you improve the design of your experiment in part a?*

Breaking items into smaller parts has implications for scoring—should the new question section and initial question section be scored individually or together? When the experimental design and data table in the flower item were separated into two parts, the standards for obtaining a score of 3 became more demanding. We also decided that for a level 3 answer, the data table provided must make sense with the experimental design proposed, which adds rigor to the assessment but also complicates scoring.

When an educator is interested in students' thinking about an open-ended item that can be approached in several ways, it may help to develop codes for the nature of students' approaches to the items in addition to scores for the level of their response. Codes were helpful in analyzing students' responses to the flower item, but not as necessary for the antibiotic item, in which a good response always included the main concept of antibiotic resistance and a weighing of this serious consequence against the less serious side effect of a stomachache.

The results of this study are preliminary. They are based on relatively small pilot tests of different student groups, with different teachers in different school districts. Student groups were not controlled or selected and may vary in many ways. In addition, data about student opportunity to learn is restricted to teachers' self reports of which course units and activities they taught. This study also involved only posttests, and does not indicate anything about student

learning over the year. As part of the larger study currently underway, psychometric analysis of pilot test data from 2002–2003 has established item fit of all items and separation reliability of the full item bank. During the 2003-2004 academic year, a pretest/posttest evaluation of student learning is being conducted at five sites. This study includes data about student populations and about teachers' classroom experience. It also includes more in-depth surveys of teachers' classroom implementation of the instructional materials.

We end with a note of caution. Teachers might have heard that multiple choice items are hard to write and easy to score and that essay items are easy to write and hard to score. Effective essay items are more difficult to write than this suggests. Open-ended items by their very nature are subject to a wide variety of interpretations by students and may elicit reasonable answers that have little to do with the science concept that the test developer intended to assess. We have found that extended response items often require several cycles of revision in order to become effective items.

Acknowledgements

We would like to thank Jenny Garcia de Osuna and Leon Goe for their contributions to scoring items for this study. We would also like to thank Manisha Hariani, Ann Kindfield, and participating teachers for their review of assessment items.

This project was supported, in part, by the National Science Foundation (ESI–9553877). Opinions expressed are those of the authors and not necessarily those of the Foundation.

References

National Research Council (NRC). (1996). *National science education standards*. Washington, DC: National Science Board.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., and Glaser, R., editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Science Board.

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

SEPUP. (2001). *Science and life issues*. Ronkonkoma, NY: Lab-Aids, Inc.

Siegel, M.A., Nagle, B., & Barter, A. (2004). Development of an assessment instrument for middle school life science: Design considerations and consequences related to scoring student learning with rubrics. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Siegel, M.A., Hynds, P., Siciliano, M., & Nagle, B. (in press). Using rubrics successfully: How to foster meaningful learning and self-assessment. *PEERS (Practical Experience and Educational Research) Matter*. Washington DC: Joint publication of NSTA and NARST.

Stiggins, R.J. (2001). *Student-involved classroom assessment*. Upper Saddle River, NJ: Merrill Prentice Hall.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system.□ *Applied Measurement in Education,* 13 (2), 181-208.

Scoring Guide Appendix

Scoring Guide: Designing and Conducting Investigations (DCI) Variable (SEPUP, 2001)

Score	Recording Design or Procedure What to look for: Response states general approach for the investigation, or procedures are described completely, accurately and safely.	Organizing Data What to look for: Response accurately records and logically displays data.	Analyzing and Interpreting Data What to look for: Response accurately summarizes data; detects patterns and trends; and draws valid conclusions based on the data used.
4	Accomplishes Level 3 AND goes beyond in some significant way, e.g. identifies alternate procedures, suggests improved materials, or relates clearly to scientific principles and approaches.	Accomplishes Level 3 AND goes beyond in significant way, e.g. innovation in the organization or display of data.	Accomplishes Level 3 AND goes beyond in significant way, e.g. explaining unexpected results, judging the value of investigation, suggesting additional relevant investigation.
3	Appropriate design with reproducible procedure if required.	Logically reflects complete and accurate data.	Analyzes and interprets data correctly and completely; conclusion is compatible with data analysis.
2	Incomplete design/procedure AND/OR significant errors.	Reports data logically BUT records are incomplete.	Notes patterns or trends but does so incompletely.
1	Incorrect or inappropriate design/procedure.	Reports data BUT records are illogical and/or contain major errors in the data.	Attempts an interpretation, but ideas are illogical OR show a lack of understanding.
0	Missing, illegible, or irrelevant design/procedure.	Missing, illegible, or no record of data included.	Missing, illegible, or no analysis or interpretation of data included.
x	Student had no opportunity to respond.		

Scoring Guide: Understanding Concepts (UC) Variable (SEPUP, 2001)

Score	Applying Relevant Content What to look for: Response uses relevant scientific information in situations such as solving problems or resolving issues.
4	Accomplishes Level 3 AND extends beyond in some significant way, e.g. relating to one's own life or to scientific concepts or themes.
3	Accurately and completely uses scientific information to solve problem or resolve issue.
2	Shows an attempt to use scientific information BUT the explanation is incomplete; also may have minor errors.
1	Uses scientific information incorrectly and/or provides incorrect scientific information; OR provides correct scientific information BUT does not use it.
0	Missing, illegible, or is irrelevant or off topic.
x	Student had no opportunity to respond.

Scoring Guide: Evidence and Trade-offs (ET) Variable (SEPUP, 2001)

Score	Recognizing Relevant Evidence What to look for: States correct and relevant evidence (such as facts, data and observations).	Using Evidence to Make Trade-offs What to look for: Response uses evidence to compare multiple options in order to make a choice.
4	Response accomplishes level 3, AND goes beyond in some significant way, e.g. questioning or justifying the source, validity, and/or quantity of the evidence.	Accomplishes Level 3 AND goes beyond in some significant way, e.g., suggesting additional evidence beyond the activity that would influence choices in specific ways, or questioning the source, validity, and/or quantity of the evidence and explaining how it influences choice.
3	Identifies key evidence with the appropriate number of facts, data and observations.	Uses relevant and accurate evidence to compare multiple options, and makes a choice based on the comparison.
2	Missing key evidence OR insufficient number of facts, data and observations.	Compares options using evidence BUT reasons or choices are incomplete and/or part of the evidence is missing; OR only one complete and accurate perspective has been provided.
1	Missing key evidence AND insufficient number of facts, data, and observations; OR states opinion as facts.	States at least one option BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	Missing; illegible, or offers no evidence.	Missing, illegible, or completely lacks reasons and evidence.
x	Student had no opportunity to respond.	