Development of an Assessment Instrument for Middle School Life Science: Design Considerations and Consequences Related to Scoring Student Learning with Rubrics

Marcelle A. Siegel and Barbara Nagle
Science Education for Public Understanding Program (SEPUP)
Lawrence Hall of Science
University of California at Berkeley
Berkeley, CA 94720-5200
http://sepuplhs.org
mcgull@berkeley.edu

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

A paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 12-16, 2004.

Abstract
In this paper, we explore the methodological challenges faced in fine tuning rubrics for scoring student learning in a middle school life science course, *Science and Life Issues*. Items included nine extended written response questions that were administered along with short answers and multiple-choice items as part of a larger study of conceptual understanding and reasoning in life science. General rubrics were developed previously. Using a process called moderation, we developed task-specific rubrics to provide more detail than the general rubrics. Data from a pretest/posttest with six school districts was collected. In this paper, we describe four of the task-specific rubrics that were developed and analyze their effectiveness based on the student results. These rubrics were found to be valid and effective at measuring students' progress, with the first two slightly more so than the second two. We discuss the decisions that were made during development of the rubrics (such as whether to add half levels between the whole scoring levels) and the consequences of these decisions.

Introduction

The nation's emphasis on science assessment has changed from assessing what is easily measured to assessing what is most highly valued, namely rich well-structured scientific knowledge and scientific reasoning (NRC, 1996). This change in emphasis requires new types of assessments and new ways of scoring assessments. The rubric, or scoring guide, is one way of scoring reform-based assessments.

In this paper, we explore the methodological challenges faced in fine tuning rubrics for scoring student learning in a middle school life science course, *Science and Life Issues* ("SALI", SEPUP, 2001). During a previous research project (e.g., Roberts, Wilson, & Draney, 1997), rubrics were developed that were analytic (vs. holistic) and general (vs. task specific). Then, as the SALI course was developed, these rubrics were refined for use by students and teachers in the SALI course. In current research, we revised the general rubrics to create task-specific rubrics, with the goal of improving the items and the scoring. Items included nine extended written response questions that were administered along with short answers and multiple-choice items as part of a larger study of conceptual understanding and process skills in life science. The learning goals and the process of item development are described in another paper (Nagle, Siegel, & Barter, 2004). Recommended ways for teachers to foster learning through rubric assessments are also described elsewhere (Siegel, Hynds, Siciliano, & Nagle, in press).

We developed task-specific rubrics using a process called moderation. In this paper, we focus on two questions:

1) What task-specific rubrics were developed and how effective were they in terms of assessing the intended content and student learning?
2) What design decisions were made and why? What were the consequences of the decisions?

Theoretical Framework Related to Assessment

Assessment researchers have categorized different types of rubrics. *Holistic* rubrics measure a performance or product along an overall trait, whereas *analytic* rubrics divide a performance or product into several traits or dimensions that are measured separately (Arter & McTighe, 2001). Another distinction is that *general* rubrics can be used to measure a set of tasks or products, whereas *task-specific* rubrics are used to measure one particular task or product (Moskal, 2000). General rubrics can be used to help students understand the central characteristics of quality work. Task-specific rubrics are especially helpful during scoring; they reduce the cognitive load of a scorer because specific standards have been set (Arter & McTighe, 2001). One danger of task-specific rubrics is that they become too tailored to specific tasks and less related to constructs, thereby limiting generalizability (Messick, 1994).

In collaboration with the Berkeley Evaluation and Assessment Research center at UC Berkeley's Graduate School of Education, the Science Education for Public Understanding Program (SEPUP) at the Lawrence Hall of Science developed an embedded, authentic assessment system as an integrated component of our first full-year science course (SEPUP, 1996; Roberts, Wilson, & Draney, 1997). The basic assessment system included three components:

- the five variables that define the key domains in which students are expected to make progress during the year,
- the actual assessment tasks, and
- the rubrics used to evaluate student performance on the tasks.

Each variable has an associated rubric that provides criteria for different levels of student performance. The development of the assessment system was based on four principles, so that:

1) it is grounded in a developmental perspective of student learning over the course of a school year,
2) assessment matches instructional goals,
3) it maintains standards of fairness, ensuring validity, reliability, generalizability, and equity,
4) teachers can use the assessment evidence to guide the learning process (Wilson & Sloane, 2000).

Introduction to the Assessment System

Our assessment system measures five constructs, called "variables." Variables include different types of content and process learning that are central to the instructional materials. Variables include: Designing and Conducting Investigations, Evidence and Tradeoffs, Understanding Concepts, Communicating Scientific Information, and Group Interaction. These five variables represent student learning in terms of the core concepts of SEPUP courses that emphasize decision making about societal issues. Three variables were the focus of the current study (summarized in Figure 1).

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.

Figure 1. Three Variables in the Current Study

Each variable has an associated scoring rubric that sets forth the expected levels of performance for students. Levels 0-4 are similar across rubrics in that: 0 indicates an off-topic or missing response; 1 indicates an attempt at a response; 2 is partially correct; 3 is complete and correct; 4 goes above and beyond the expected complete response for the question.

Method

Student Participants

Over 600 seventh-grade students from six school districts (from four states) participated. Three of these districts have student populations with a significant number of English learners and under-represented groups in science. The data were reduced to 275 by randomly selecting student papers to score. For matched analyses, only students with complete tests (specifically: both parts of the test and both pretest and posttest) were included. Because there were multiple forms of the test with mostly different items, only students who took an item on both the pretest and posttest were included. The final number for the matched analyses thus varied depending on student/form/item, but was typically less than 100.

Curriculum

SALI is a year-long program developed with funding from the National Science Foundation. Seven units address: the scientific process, body systems, microbiology, genetics, ecology, evolution, and bioengineering. A Student Book provides laboratory experiences, investigations, and readings; a kit provides materials for hands-on investigations; and a 1,000-page Teachers Guide provides suggested teaching approaches, scientific background, information about students' ideas and possible responses, and suggestions for modifying or enhancing lessons for different student groups. Course materials also include an embedded assessment system developed in collaboration with the Berkeley Evaluation and Assessment Research center. The SEPUP Assessment System is presented as an exemplary model of measurement in the National Research Council's report, *Knowing What Students Know* (NRC, 2001).

Instruments

The extended response items for this study are part of a test designed to correlate with the SALI course and the national standards in life science, inquiry, and science in personal and social perspectives. Multiple forms of the test containing mostly different items were created in order to test all the items. Items address each SALI unit and have been piloted for two years to create a valid and reliable pretest/posttest.

Items were first piloted in a small number of classrooms using the SALI curriculum. Researchers and curriculum developers reviewed the results and revised or eliminated items that were not effective at eliciting seventh graders' knowledge. In the second year of item development, over 500 students completed posttests comprised of 162 items (each student spent two periods on a subset of items). Psychometric analyses were conducted looking

at item fit, item discrimination, and separation reliability. Of the 162 items, only one item did not fit the model. Item discrimination statistics found a reasonable range of item discrimination, and the separation reliability of the items was 0.99. Maps of items by item difficulty were developed in order to ensure that there was an appropriate spread of items from easy to difficult for each type of item.

Testing Procedure

Student answers to nine extended items (and additional items not discussed here) were collected during a pretest at the beginning of school year 2002-03 and a posttest at the end. Teachers administered the tests according to written guidelines. Students spent one class period answering the extended items (and another class period on the multiple choice and short answer items). There were several forms of the test. Students were randomly given a form of the test. Because most of the items did not overlap on the forms, it was possible for a student to not have the same items on the pretest and posttest. This was helpful for pilot test purposes so there would be time to test all of the items developed.

Rubric Development

The task-specific rubrics were developed through a process called assessment ***moderation*** in which teachers/researchers use rubrics to score a set of student papers and then discuss their reactions to the differences in student scores and the rubric itself in an effort to reach consensus (e.g., Roberts, Sloane, & Wilson, 1996). Moderation has been described as a time-consuming process that can greatly enhance reliability (Yancey, 1999).

Throughout 2002-2003, the project team met four times to moderate different sets of items. We selected student responses that were intended to be representative of the range of student performances for that item, and ensured that there were enough at the higher levels to discuss. (The responses at the lowest levels do not usually provide enough content to discuss.) Each researcher then scored the selected responses using the general rubric. During a three-hour moderation meeting, we discussed the scores and made decisions about what scoring level each response should be. In addition, we made greater decisions about the scoring levels (e.g., why does this response represent a 2?), and the number of scoring levels (e.g., are additional half steps needed?). Then, the scorer for that item would use the results of the moderation session to draft a detailed scoring rubric that was task-specific for the item. Sometimes s/he scored additional items in order to make a final decision on one of the issues that arose during the moderation meeting. All researchers then read the drafts and provided comments if appropriate. The final task-specific rubric was used to score the remainder of the items. The final task-specific rubrics were also evaluated by an external consultant with expertise in science assessment.

Analysis of Student Data

Two graduate student researchers scored the extended items using the task-specific rubrics. They scored all tests for certain items. Random checks of scoring were conducted (but double scoring and tests of inter-rater reliability were not conducted due to cost restraints).

Student data was analyzed using basic statistical tests with SPSS software and through brief qualitative summaries of sample responses.

Results

Fifteen detailed rubrics were developed as a result of the study. Several of the extended items were scored with more than one rubric. Four rubrics for two items will be discussed here. Results will be provided for the task-specific rubrics developed and the effectiveness of the rubric/item. The final results section describes some of the design decisions—defining the scoring criteria and adding half levels—and their consequences.

Results for the Microbiology Item

Task-Specific Rubrics

An item was designed to be scored with the UC and ET rubrics. The item was about human use of and bacterial resistance to antibiotics (see Figure 2). The item was designed to find out students' understanding of what they had learned in the SALI course: that antibiotic drugs are specific to bacteria, that some bacteria are more resistant than others to the drugs, and that these would be the ones left to reproduce if the full course of antibiotics is not taken. Students' reasoning was assessed in the context of a decision about whether to stop taking antibiotics due to a mild side effect. The item (along with others) aligns with Content Standards C and F from the National Science Education Standards (NSES) for grades 5-8 (Life Science—Structure and Function, and Science in Personal and Social Perspectives—Risks and Benefits).

Based on the UC variable, we developed the task-specific rubric shown in Table 1. Notice that the task-specific rubric is based on the same guidelines as the general rubric, but describes specific science content related to the particular item. Also, we added more scoring levels—half levels—to the new rubric.

> **Item:** 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.*
> (Research Item for Micro-life Unit of SALI.)

Figure 2. Microbiology Item

Table 1. General and Final Task-Specific UC Rubrics for Microbiology Item

	Before the Study	*After the Study*
Level	**General Rubric**	**Task-Specific Rubric**
3.5/4*	Accomplishes Level 3 AND extends beyond in some significant way, e. g. relating to one's own life or to scientific concepts or themes.	Accomplishes level 3 and goes beyond in some significant way. For example: Antibiotics should continue to be taken, because otherwise it would be possible that not all bacteria had been killed. The remaining drug-resistant bacteria that had not been killed would multiply. It would be more difficult to eradicate these bacteria, because of their resistance to the original antibiotics. Other types of antibiotics would be required to kill these bacteria.
3	Accurately and completely uses scientific information to solve problem or resolve issue.	Accurately and completely presents the statement that the antibiotics should continue to be taken, because otherwise it would be possible that not all bacteria had been killed. The bacteria left in Rita's body would be the "stronger" bacteria. These "stronger" bacteria would multiply and antibiotics would not work on these bacteria.
2.5		Accomplishes level 2 and beyond with more complete explanation. Refers to strength of bacteria to cause damage if antibiotics are stopped.
2	Shows an attempt to use scientific information BUT the explanation is incomplete; also may have minor errors.	Shows an attempt to provide the explanation above, but the explanation is incomplete or may have minor errors. For example: the student does not mention that the antibiotics would not work on the "stronger" bacteria left in Rita's body.
1.5		Accomplished level 1 but provides some explanation.
1	Uses scientific information incorrectly and/ or provides incorrect scientific information; OR provides correct scientific information, BUT does not use it.	(Provides incorrect scientific information and/or uses scientific information incorrectly.)
0	Missing, illegible, or is irrelevant or off topic.	(No answer or irrelevant answer.)

*General is 4; task-specific is 3.5.

For a complete and correct score of level 3, students needed to refer to the "strength" or resistance of bacteria. We accepted references to "stronger" bacteria from the seventh grade students, rather than a more complete explanation, such as "There is variation in the bacterial population, and some bacteria are more resistant to the antibiotic."

Based on the ET variable, the task-specific rubric shown in Table 2 was developed.

Table 2. General and Final Task-Specific ET Rubrics for Microbiology Item

	Before the Study	*After the Study*
Level	**General Rubric**	**Task-Specific Rubric**
3.5/4*	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.	Includes everything in level 3 plus the phrase "antibiotics-resistant". Explanation that goes beyond level 3.
3	Uses relevant and accurate evidence to compare multiple options, and makes a choice based on the comparison.	Uses relevant and accurate evidence to compare multiple options, and makes a choice based on the comparison. Mentions trade-offs or advantages vs. disadvantages. For example: Continuing the antibiotics would upset Rita's stomach. Discontinuing the antibiotics after only three days would create the risk of the "stronger" bacteria multiplying.
2.5		Accomplishes level 2 with at least pieces of evidence.
2	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.	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. For example: "The trade-off is that she still will have the stomach problem, but if it is really bothering her, she should see her doctor."
1.5		Accomplishes level 1, and attempts to provide alternative reasoning other than subjective or inaccurate statement.
1	States at least one option BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.	States at least one option, but only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	Missing, illegible, or completely lacks reasons and evidence.	No answer or irrelevant answer.

*General is 4; task-specific is 3.5.

Effectiveness

The task-specific rubrics generated for the microbiology item appeared to meet criteria for content and construct validity (see Moskal & Leydens, 2000 for an application of validity issues to scoring rubrics and not just items). Content validity of the rubric was addressed in that the new rubrics did not address extraneous content, they addressed all aspects of the intended content, and they did not exclude any important content from being measured. The ET rubric, for instance, provides specific pieces of evidence covering the advantages and

disadvantages of continuing or discontinuing the antibiotic regimen. As far as being comprehensive, there were some student answers that were of course not mentioned in the task-specific rubric, but none that we considered essential. For example, one student idea not included in the rubric was a suggestion to continue using the antibiotic, but eliminate the upset stomach. Some students suggested using "Pepto Bismol" and some recommended trying a different antibiotic. Also, construct validity for the two task-specific rubrics were met in that the important facets of the intended construct, laid out by the general rubrics, were met, and no extraneous evaluation criteria were added to the constructs.

The task-specific rubrics also seemed to be effective for each level of the rubrics. The ET rubric, for example had many responses matching the criteria in the new rubric. One student sample for each level of ET is provided below:

Score of 3.5

"Well, if she stops the treatment she would only be killing the least resistant bacteria. There are three categories of bacterial infection: least resistant, resistant, most resistant.

When you start the treatment, it slowly kills the least resistant and works its way up to the most resistant.

But if you do stop taking the antibiotics, the ones that are left behind are the most resistant bacteria. Then they reproduce and cause you much more problems. Then you will take longer to kill. This is only a temporary solution to feeling better.

If you don't you can get over the infection and not have to take any more medicine that will hurt your stomach."

Score of 3

"Rita should continue to take her antibiotics.

Continue taking them:
Advantages:
-All of the bacteria will be killed so they will not be able to reproduce
-won't re-appear
-will feel better in the end
Disadvantages:
-upset stomach
-she already feels better so it's a pain to take it.

Stop taking the:
Advantages:
-no upset stomach
-don't have to take it
Disadvantage:
-sickness will come back

"Continuing taking them" is the smarter choice for Rita. The advantages outweigh the disadvantages."

Score of 2.5:

"Rita should definitely continue taking the antibiotics. If she stops, the remaining bacteria will reproduce again and the sickness will start over again. She may have stomach problems though and may not like it, but at least the sickness will be going away. She should just take the pills and get it over with! My final recommendation to Rita is to continue taking the antibiotics until she is allowed to stop."

Score of 2:
"If she stops taking antibiotics, then she will get sick again. If she keeps taking them, she will feel better but there will be more side effects."

Score of 1.5:
Yes, so that Rita will soon feel better and if her stomach hurts she could take another medicine to help her stomach.

Score of 1:
"She should stop and maybe try something else."

With the student data, the two task-specific rubrics also provided a range of student response levels. One would expect more low scores and blank answer sheets on the pretest and a range of response levels on the posttest. Such a range was found as shown in Table 3.

Table 3. Frequency of Student Scores for Microbiology Item

Level	Frequency	
	Pretest	Posttest
UC		
0	4	8
1	8	8
1.5	0	7
2	20	13
2.5	0	26
3	1	15
3.5	0	7
Total	33	84
ET		
0	2	8
1	10	9
1.5	0	3
2	18	19
2.5	0	26
3	3	19
3.5	0	0
Total	33	84

The new rubrics also seemed effective in that they captured student learning over time, as one would expect with a pretest/posttest design. On average, scores were 1.29 on UC and 1.47 on ET on the pretest and gained about one full point on the posttest. Table 4 displays the results for the matched group of 84 students on the pretest/posttest. Improvement was significant for both UC and ET variables; for UC: $t(32)=7.21$, $p<.000$; for ET: $t(32) =6.64$, $p <.000$.

Table 4. Student Results on Microbiology Item

	Pretest		Posttest	
	UC	ET	UC	ET
Mean	1.29	1.47	2.35	2.32
Maximum	3	3	3.5	3
Standard Deviation	.72	.66	.97	.89

Results for the Medicine Item

Task-Specific Rubrics

An item on analyzing data from a clinical trial of a cough medicine and making a decision about using it on humans (see Figure 2) was designed to be scored with the DCI and ET rubrics. The item was designed to assess students' understanding of experimental design and controlling variables for a clinical trial and their interpretation of a data table. It was also designed to assess how students used evidence from the data table to form a decision about whether to sell the new medicine. The item (along with others) aligns to Content Standards A, C, and F, from the NSES for grades 5-8 (Inquiry, Life Science—Structure and Function, and Science in Personal and Social Perspectives—Risks and Benefits).

Based on the DCI variable about analyzing and interpreting data, we developed the task-specific rubric shown in Table 1. Again, the task-specific rubric is based on the same guidelines as the general rubric, but describes specific science content related to the particular item. Also, we added extra scoring levels—half levels—to the new rubric.

Item: Scientists have performed a trial of a new cough medicine. They divided a group of patients with a bad cough into two similar groups. Each group included males and females and people of different ages. Group A received cough syrup. Group B received plain syrup that did not contain any cough medicine. Every day for four days, the scientists interviewed the patients to find out whether their coughs were as frequent and as serious. They also asked the patients if they had any new health problems while taking the medicine. The following table summarizes the data.

Group	Total number of patients	Number who feel better	Number who feel the same	Number who feel worse	Number with side effects (dizziness and stomach upsets)
A (cough syrup)	50	40	5	5	10
B (plain syrup)	50	20	25	5	2

A. Analyze the data to form a conclusion about how well the medicine works. Explain your answer.

B. Should the cough medicine be sold? Be sure to include the advantages and disadvantages of both choosing to sell and choosing not to sell the medicine and to explain the trade-offs of your final decision.

(Research Item for the Studying People Scientifically Unit of SALI.)

Figure 3. Medicine Item

Table 5. General and Final Task-Specific DCI Rubrics for Medicine Item

	Before the Study	***After the Study***
Level	**General Rubric**	**Task-Specific Rubric**
4	Accomplishes Level 3 AND goes beyond in significant way, e.g. explaining unexpected results, judging the value of investigation, suggesting additional relevant investigation.	(Same)
3	Analyzes and interprets data correctly and completely; conclusion is compatible with data analysis.	-Addresses all or most of the columns in table (quantitatively or qualitatively). quantitative = "half, three fourths" not words like "little, or some" -Compares most of A and B (reference to a control is best). -Uses all or most of the data in table.
2.5		Uses at least two or more pieces of data from table or gives thorough qualitative descriptions of A but does not compare it with B, or compares A and B with at least one piece of data for each.
2	Notes patterns or trends but does so incompletely.	-Gives quantitative/qualitative interpretation for part of the table -Uses some of the data presented in the table (at least 2 items) or qualitatively explains two or more trends
1.5		Uses one piece of data from table or gives qualitative summary of at least two trends (e.g., "twice as much").
1	Attempts an interpretation, but ideas are illogical OR show a lack of understanding.	-Wrong interpretation of part of the table (e.g., "60 feel better so it works"). -Only one trend is mentioned (e.g., side effects) that doesn't really answer the question. -No use of data presented in the table for analysis. -General descriptive statements or trends mentioned.
0	Missing, illegible, or no analysis or interpretation of data included.	-Really illogical -No explanation given -No analysis of any sort -Only conclusion given (e.g. "the cough syrup") -States conclusion that doesn't answer question

Based on the ET variable (see Table 2), the task-specific rubric shown in Table 6 was developed.

Table 6. Final Task-Specific ET Rubric for Medicine Item
(For general ET rubric, see Table 2)

Level	Task-Specific Rubric
4	(Same as general rubric)
3	Discuss two options with reference to two quantitative pieces of data for each option (e.g., "yes because…no because…").
2.5	Discuss two options with reference to one quantitative piece of data for each option (e.g., "yes because…no because…").
2	-One option plus three or more quantitative pieces of data -Complete qualitative summary of both sides -Two options with at least one qualitative reason for each
1.5	-One option plus two quantitative pieces of data (e.g., "yes because…no because …") -Complete qualitative summary of one side -Partial qualitative summary of both sides
1	-One option, one piece of data -Uses poor evidence (e.g., "60 got better") -States position(s) without conclusive evidence for either side other than general subjective reasons -States option with misinterpretation of table
0	-Really illogical -Just says "yes"

Effectiveness

The task-specific rubrics developed for the medicine item appeared to meet criteria for content and construct validity. Content validity of the rubric was addressed in that the new rubrics did not address extraneous content, they addressed all aspects of the intended content, and they did not exclude any content from being measured. Construct validity for the two task-specific rubrics were met in that the important facets of the intended construct, laid out by the general rubrics, were met, and no extraneous criteria were added to the constructs.

This rubric was difficult because of the multitude of possible student answers based on the complex data table. One aspect of the rubric that was improved, but perhaps could be improved further has to do with the level of detail vs. level of abstraction for key differences in responses. After moderation, the rubric for

level 2 was: 'the student gives a quantitative interpretation for part of the table and interprets at least 2 of the items in the table.' It was detailed but perhaps did not get at the important content. After review and input from the external consultant, level 2 was distinguished from level 3 further, by saying that responses at level 2 do not compare the data from the two subject groups presented in the table. However, responses at level 3 do provide a comparison and may refer to the "control group." An example of a level 2 response was: "I conclude the medicine was a success even though some patients did not feel better and some felt worse. I conclude this because 80% of the patients in group A felt better and only 20% felt side effects." Notice that the response does not refer to patients in Group B.

Another way the task-specific rubrics for the medicine item were deemed marginally effective was based on frequency data. A small range of student response levels was found for both of the rubrics for the medicine item, as shown in Table 7. In the matched dataset, there were no responses at the 1.5 or 2.5 levels of the DCI rubric, and few for the ET rubric. However, in the unmatched dataset of 275, there were more of these, with the most for level 1.5 of ET: 45, or 16%.

Table 7. Frequency of Student Scores for Medicine Item

Level	Frequency	
	Pretest	Posttest
DCI		
0	18	15
1	22	21
1.5	0	0
2	10	31
2.5	0	0
3	1	17
3.5	0	
Total	51	84
ET		
0	11	15
1	34	46
1.5	0	0
2	6	3
2.5	0	4
3	0	0
3.5	0	0
Total	51	84

The new rubrics also were evaluated based on how well they captured student learning over time. On average, scores were below 1 on DCI and ET on the pretest. The DCI posttest mean rose to 1.47, while the ET posttest mean remained about the same. Table 8 displays the results for the matched group of 84 students on the pretest/posttest. Improvement was significant for the DCI variable, $t(50)=3.27$, $p=.002$. The slight decrease for the ET variable was not significant, $t(50)=.0056$, *ns*.

Table 8. Student Results on Medicine Item

	Pretest		Posttest	
	DCI	ET	DCI	ET
Mean	.88	.84	1.47	.75
Maximum	3	2	3	2
Standard Deviation	0.79	.47	1.08	.48

The medicine item appears to be an example of an item that needs further revision, not necessarily the rubric. Reviewing student responses showed that they had trouble with reading and interpreting the item. Often students did not write a complete answer. Students neglected many columns in the table, especially the side effects column. Teachers reported that students were confused by the terms "new health problems," "side effects" and the idea of how well the medicine worked. Perhaps, this item is too complex to demonstrate what seventh grade students know. We have been revising the item for language – to make references more parallel, to shorten it, and to simplify the language. And we have been revising the item's data table to decrease the amount of data included. Then we will retest to see if the new item is more effective.

Design Decisions and Consequences

In this section, we take stock of design decisions that were made during rubric development and the implications of those decisions.

One issue was how to define the criteria for scoring levels. Would the choice of scoring certain answers a 1, for instance, instead of a 2 affect our study? One item involved designing a procedure for an investigation of the height of students at school. Through moderation, we decided that it was necessary to have a reproducible procedure, but it was okay to leave out some specifics to attain a level 3. For a level 2, students could provide a procedure that was not reproducible, or say that they would do a survey and ask people their heights, but not explain the procedure more than that. For a level 1, the procedure was incorrect or inappropriate. Results showed that the average for this item was .93 (sd=.10) on the pretest and 1.15 (sd=1.2) on the posttest. This was an item that showed a very small positive increase from the pretest to posttest (and not statistically significant). Thus, one could ask if the scoring scale was sensitive to potential knowledge growth. If we had been less strict about criteria for levels 2 and 3, there might have been a greater range of responses and more responses at level 2 on the posttest. However, the overall scoring for that item would have been less difficult (easier to get a high score because of lower standards). A less difficult scoring scale might have meant we would see a greater range of responses, but it would have involved a tradeoff regarding high standards. For this particular example, we would argue that the scoring criteria were sound. The poor performance could be explained by looking more at the item itself, looking at other items on the same topic of designing a scientific investigation, or

student variables (such as, students did not spend enough time writing answers to the question, as often happens in seventh grade).

Another issue that arose was whether to create half levels. We ended up creating half levels for certain questions for which the extra levels helped to distinguish between different quality performances, but not for other questions. For example, for the medicine item, we created a level 1.5 and a level 2.5. This item contained many pieces of information in the data table for students to analyze. The half levels represented responses that went beyond level 1 or level 2 in amount of description and interpretation of the data, but without making a conceptual leap to the next whole level (such as level 3 that requires a comparison of the control and experimental subject groups). One possible consequence of adding more levels to a rubric is that it might be more difficult for the scorer to judge using so many criteria. On the other hand, the scorer might have an easier time because the specific criteria make sense for the item. Further work could be done to investigate this with our data.

Another possible consequence of adding half levels to a rubric is that you might detect more change between the pretest and posttest. For example, if students received a 2 on the pretest and posttest, if you add half levels to the rubric, perhaps they would score a 2 on the pretest and a 2.5 on a posttest. We began to investigate this question by recoding the half level data to whole levels for the microbiology item. Half levels were changed to the level below (2.5 to 2) because that is how the task-specific rubrics were designed, based on the general rubrics. Mean scores for the recoded (no half levels) data were the same on the pretest and decreased on the posttest. The gains from pretest to posttest were very significant for the half level data, and were also significant for the recoded data, but the t values were less (t=5.4 for UC and 5.5 for ET). For this item, adding the half levels during the development process seemed to aid scoring, but did not radically enhance our ability to capture learning gains.

General Discussion

This study has resulted in the development of fifteen task-specific rubrics, four of which have been discussed here. The rubrics and items assess concepts that relate to the seventh grade curriculum and NSES content standards for life science, inquiry, and science in personal and societal perspectives. The task-specific rubrics for the microbiology item appeared to be valid for purpose and effective at measuring targeted aspects of students' understanding and decision making. The task-specific rubrics for the medicine item were designed to assess certain aspects of students' analysis of data and weighing of evidence. These rubrics appeared to be effective, but less so, and the item itself is currently under revision.

The technical results of this study are preliminary as they are based on relatively small pilot tests. As part of the larger study currently underway, psychometric analyses of pilot test data from 2002–2003 has established several types of reliability and validity of the instrument. During the 2003-2004 academic year, a pretest/posttest evaluation of student learning is being conducted at five sites.

This evaluation includes data about student populations and about teachers' classroom experience. It also includes more in-depth surveys of teachers' use of the instructional materials in the classroom.

The design of our task-specific rubrics was based on several general rubrics that assess constructs called variables (understanding of concepts, designing investigations, using evidence to make decisions, etc.). Because of this design, the danger of creating rubrics that are too specific to items and that lose their connection to constructs of learning was perhaps avoided. The task-specific rubrics were specific to items while still reflecting the variables at levels 0,1,2,3 and 4. Another possible strategy for avoiding this danger, recommended by Messick (1994), is to aim for scoring rubrics that are neither too specific to the task nor generic to the construct, but are in some middle ground that reflects the classes of tasks that the construct will generalize to.

We have explained a few design considerations and consequences of developing scoring rubrics for a life science assessment instrument. We have found that developing effective rubrics takes many cycles of revision and many types of analyses. We hope these kinds of analyses will be of use to others who are developing assessment systems. The methodological issues we faced provide starting points for discussion of ways to justify evidence-based claims in education.

Acknowledgements

We would especially like to thank Leon Goe, and also Jennifer Garcia de Osuna, Tony Lin, and Thanh Le for their contributions to moderating and scoring items for pilot and current research. We are grateful to Ann Kindfield, Manisha Hariani, 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

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance.* Thousand Oaks, CA: Corwin Press.

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

Moskal, B.M. (2000). Scoring rubrics: What, when, and how? *Practical Assessment, Research & Evaluation* 7 (3). Retrieved July 27, 2003 from http://edresearch.org/pare/getvn.asp?v=7&n=3.

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation, 7* (10). Retrieved July 30, 2003 from http://edresearch.org/pare/getvn.asp?v=7&n=10.

Nagle, B., Siegel, M.A., & Barter, A. (2004). Evolution of life science assessments for middle school. Paper presented at the 2004 annual meeting of the National Association for Research on Science Teaching, Vancouver, CN.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaser, R., (Eds.). 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.

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

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

SEPUP. (1996). *Issues evidence and you.* Ronkonkoma, NY: Lab-Aids, Inc.

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.

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

Yancey, K.B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50,* 483-503.