SEPUP Course I, *Issues, Evidence and You*: Achievement Evidence from the Pilot Implementation

Mark Wilson, Kathryn Sloane, Lily Roberts & Robin Henke
University of California, Berkeley

June 1995

Abstract

The Science Education for Public Understanding Project (SEPUP) has developed an issue-oriented science curriculum for the middle/junior high school entitled *Course I: Issues, Evidence, and You*. This paper reports an evaluation of its initial implementation year. The evaluation used a traditional pretest-posttest design to compare change over the year between students in SEPUP classes and a group of comparison classes on the "Evidence and Trade-offs" variable. The quantitative comparison is augmented by a qualitative map of the map which allows one to make educational interpretations of the quantitative results. The SEPUP students were found to have made gains which are both statistically and educationally significant. Although the comparison students made gains during the year, these were not statistically significant.

Acknowledgments: This project has been supported by NSF Grant No. MDR9252906, and also by other support through SEPUP at the Lawrence Hall of Science. We would like to thank the teachers and students involved in the Course I pilot year for their cooperation and enthusiasm. We would like to thank Chris White and Eric Crane for data management and data analysis for this report.

SEPUP Course I, *Issues, Evidence and You*: Achievement Evidence from the Pilot Implementation

The Science Education for Public Understanding Project (SEPUP) at the Lawrence Hall of Science has developed a new year-long course in issue-oriented science for the middle school and junior high grades entitled *Course I: Issues, Evidence, and You* . The development of this course, *Issues, Evidence, and You,* has included a substantial commitment to the development of an integrated system of student assessment, to be used by teachers to assess and chart student performance on the central concepts and skills that define the course (for detailed information on the assessment system see Sloane and Wilson, 1995). This document describes evidence that was collected during the initial pilot implementation during the 1993-94 school year. As the structure and instructional procedures of the course were still in a state of development throughout that year, the results reported here should be seen as indicative but not final.

Background

Course I: Issues, Evidence, and You focuses on environmentally- and socially-contextualized science content. As part of the course, students are regularly required to use scientific evidence and weigh it against other community concerns with the goal of making informed choices about relevant "real-world" issues or problems.

An innovative and integral part of SEPUP *Course I* is its instructionally-embedded assessment and evaluation system which has been developed and implemented in concert with the SEPUP curriculum. This assessment system is designed not only to improve assessment methodologies but also to reform teacher practice in the use of assessment so it can be used immediately and reflectively to improve instruction and student achievement.

The pilot curriculum program was field tested in 15 Project Development Centers (PDCs) comprised of multi-school sites in 12 states across the country during the 1993-94 school year by approximately 70 teachers and their students. The sites were not randomly selected and vary considerably on several variables including: student grade level (ranging from grade 7 to grade 9); ethnic balance (which differs considerably in the classrooms and includes students from English as well as limited- or non-English speaking backgrounds); teacher experience and familiarity with SEPUP and its associated curricula; and urbanicity of the sites (from inner city Washington D.C. to Alaska).

Purpose of the Assessment Project

The SEPUP Assessment Project is designed to apply new theories and methodologies in the field of assessment to the practice of teacher-managed, classroom-based assessment of student performance (Wilson & Adams, in press). An important aspect of this project is that the *assessment system is tied to a specific curriculum*. While compatible with other national, state, or district efforts to implement new forms of student assessment, the SEPUP assessment system is specially designed for the specific curriculum that teachers are using in their classrooms. Assessments are fully embedded in the instructional materials and are designed to be an integral part of the instructional process.

A second important feature of the Project is that the responsibility for assessing student performance, and for interpreting and using information regarding student performance, is placed *squarely in the hands of the classroom teacher*. The system provides a set of tools for teachers to use to: (a) assess student performance on central concepts and skills in the curriculum, (b) set standards of student performance, (c) track student progress over the year on the central concepts, and (d) provide feedback (to themselves, to students, to administrators, parents, or other audiences) on student progress and on the effectiveness of the curriculum materials and the classroom instruction. Initially, managing a new classroom-based system of embedded assessment demands much of the teacher. However, we believe that teachers must be recognized as the "front-line" professionals who will ultimately determine the usefulness or effectiveness of any attempt at educational reform. Empowering teachers (through the provision of the tools, procedures, and support) to collect, interpret, and present their own evidence regarding student performance is an important step in the continuing professionalization of teachers in the field of assessment.

Method

The student assessments will also be used for evaluation purposes. At this point, these data have not been fully analyzed, so this report will focus on an instrument that was used in the initial pilot year as a pre- and posttest.

Evaluation Design

Each of the PDCs were asked to nominate one of their SEPUP teachers for the evaluation design, and to locate one other teacher who could serve as a comparison. Rather than compare the SEPUP teachers, who were selected for their excellence and enthusiasm, with average teachers, we chose to make the comparison as tough as we could, and asked that the PDC director choose a teacher who might well have been chosen to be one of the SEPUP teachers--in fact, if possible, we asked that it be the “next on the list”. We did not specify

anything concerning the students of these teachers (as was the case for the SEPUP teachers), except that they be at an appropriate grade level (7 to 9).

Subjects

As is often the case in field studies, not every PDC complied completely with these instructions (two actually dropped out during the course of the trials), but we did secure test data from 26 classes, roughly split between SEPUP classes and comparison classes. This resulted in a complete data set of 830 students, with 476 from the SEPUP classes and 354 from the comparison classes. Unfortunately not all students took both pre- and posttests. In particular, 412 students from the SEPUP classes took the pretest, and 177 took the posttest, while 269 students from the comparison classes took the pretest and 98 took the posttest.

Instruments

The instruments consist of a pretest and a posttest, with common items between. The format of the items is a little unusual, so we will take a little time to describe them. Each item uses a piece of stimulus material which describes a choice that someone must make. The student is asked to choose for him/herself using a multiple choice format, then to justify that choice in a written answer. It is possible to score the multiple choice items for "correctness", and indeed we did that, and the results were consistent with those reported here, but given that the focus of Course I is a student's ability to understand a complex situation involving scientific evidence and weigh the pros and cons, we decided to concentrate our efforts on the written responses.

The pretest was developed quite early in the development of Course I, and correspondingly does not match wholly the emphases of the final course structure. There are three major variables that Course I addresses (see Sloane and Wilson, 1995 for details of these variables), and only one features strongly in the pretest. Hence, we decided to concentrate on that variable: *Evidence and Trade-offs*. This variable consists of identifying objective scientific evidence as well as evaluating the advantages and disadvantages of different possible solutions to a problem. A sample item, Item 1, related to this variable is shown in Figure 1. The responses to this item were scored using the scoring guide shown in Figure 2. Exemplars of student responses at each of the score levels are found in Figure 3. There were 5 items related to the Evidence and Trade-offs variable: All of these scoring guides were analogous[1] to that of item 1, and in the remainder of the results section, we will concentrate on those five items. All except one of the items (Item 5) were given on both pre- and posttest.

[1] Item 5 had a somewhat different scoring guide, which resulted in it not distinguishing between levels 0 and1 on the scoring guide shown in Figure 2.

Insert Figures 1, 2 and 3 about here

Analyses

Analyses were conducted using the Quest software (Adams & Khoo, 1993), which allows one to develop a map of the outcome variable. This map is based on a Rasch measurement model (Rasch 1960/80; Wright and Masters, 1982), which allows one to simultaneously place both persons and items on the same scale. We will make use of this feature in order to provide a criterion-referenced account of the Evidence and Trade-offs variable. We will also examine the students' growth over the year represented by the differences between pretest and posttest, and note the difference between growth in SEPUP classes and comparison classes. The criterion-referenced description of the variable will allow these differences to be substantively interpreted.

Results

The first map, shown in Figure 4, illustrates the relationship between the item difficulties and the students for the Evidence and Trade-offs variable. Note that in this map we have treated each student response (both pretest and posttest) as an additional piece of information helping us to map the variable--that is why there are 1660 student responses recorded here, rather than 830. To read this map, note that increasing scores are indicated by progress UP the page. The column of numbers on the left hand side indicates the units in which the map is calibrated. They are called "logits", and they are in units of log-odds: A logit is that distance on the variable that corresponds to odds of success (compared to failure) equal to *e*, the base of the natural logarithms--approximately 2.7:1. Perhaps an easier way to interpret these units is to consult Table 1, which shows the probability indicated by selected logit differences. Note that these probabilities are relative to the logit difference between a student's location and an item threshold, so that positive logit differences correspond to the student being above the item threshold. The locations of the students are indicated by the column of X's (in this case 14 students per X) in the vertical histogram. The locations of the item thresholds are indicated by the numbers on the right hand side of the vertical line symbolizing the variable. The notation is in the form n.m (e.g., 1.2): This indicates the threshold of score level m for Item n (e.g., 1.2=threshold between score levels 1 and 2 for Item 1). To interpret these locations, note that many students are located at about -1.2 logits, a little above the location of threshold 2 of Item 5 (i.e., 5.2): This indicates that we would predict that these students would have a probability of about 0.5 of getting to at least score 2 on Item 5. Note that "2.3" is about 2 logits higher than these students: This indicates that we

would predict that these same students would have a much smaller probability, about 0.12 (see Table 1), of getting to at least score 3 on item 2[2].

Insert Figure 4 and Table 1 about here

We can also use Figure 4 to make criterion-referenced interpretations of student scores on the pretest and posttest. This is because the item thresholds can be used to provide meaningful context for certain zones on the variable. This has been done in Figure 5. Here the item thresholds have been replaced by criterion zones representing the score levels used to score the student performances on the test. We can see that typical student responses range from responses that relate to the topic, but not in a way that responds to the question, to responses that give a complete answer to the question, noting more than one position on the issue, citing valid supporting information. and weighing the trade-offs involved in making a decision. Note that although no students typically make responses in either of the extreme score levels, these are included (a) because they indicate the extremes of the scores, and (b) because some students will still make individual item responses in these score levels, even though their *typical* responses are more moderate. We will use this map to make qualitative interpretations of the gains by students in SEPUP and comparison classrooms.

Insert Figure 5 about here

The performance of the students in the comparison classes, for both pretest and posttest, is shown in Figure 6. The ranges of performances at pre- and posttest are approximately the same, with a somewhat higher maximum in the posttest. This most probably reflects a maturation change over the school year, and is reflected in the small change in the means from pre- to posttest (indicated by **M** on the variable) of approximately .4 of a logit. In comparison, Figure 7 shows the same maps for the students from SEPUP classes: Here a group of students (approximately 40) have moved beyond the top of the pretest range, and, in particular, there is a small group of students responding consistently in the "complete response" category. This change at the maximum is reflected in a much larger change in the means than for the comparison students, approximately 1.4 logits. This means that the average student has changed from typically giving a response that only notes one position on an issue, to typically responding by noting at least two issues, and citing supporting evidence. As this is one of the major goals of SEPUP Course I, we see this as significant confirming evidence that

[2]Note that Item 5 does not have an entry for 5.1--this corres[onds to the point made in footnote 1.

the new curriculum is working well for a great many students. The change for the SEPUP students is clearly an educationally significant change. It is also a statistically significant change (at the α=0.05 level), as can be seen by examining Table 2, which shows the means and t-tests for the pre-post comparison for the two groups. In contrast, the change for the students in the comparison classes is not statistically significant (at the α=0.05 level). The message at the lower end is not so rosy, however, for, just as was the case among students from the comparison classes, there are students who remain at the minimum range of responses.

Insert Figures 6 and 7 and Table 2 about here

Conclusion

These results from the initial pilot year of implementation of SEPUP's *Course I: Issues, Evidence, and You* illustrate quite clearly that students in the course are doing measurably better on the core SEPUP variable than a group of their peers who have not had the benefit of the new course. That this is not just a statistical difference, but is in fact a significant and important educational change for these students, is illustrated by the map for the SEPUP students (Figure 7), where the quantitative gains may be interpreted in terms of educational goals. There are some caveats to these conclusions however: (a) as the curriculum was still under development in the 1993-94 school year, we must wait upon further years' results in order to judge the scope of the benefits that the SEPUP curriculum brings to the students, and (b) the fact that there are still SEPUP students at the minimum after a year's work on the curriculum, shows that there is still some work to be done in order to bring these benefits to *all* students.

References

Adams, R.J., & Khoo, S-K.(1993). Quest: The interactive test analysis system. Hawthorn, Australia: ACER.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Originally published 1960)

Sloane, K.S., & Wilson, M. (1995). The SEPUP assessment system. Berkeley, CA: Graduate School of Education, University of California at Berkeley.

Wilson, M., & Adams R.A. (in press). Evaluating progress with alternative assessments: A model for Chapter 1. In M.B. Kane (Ed.), *Implementing performance assessment: Promise, problems and challenges*. Hillsdale, NJ: Erlbaum.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Figure Captions

Figure 1. Item 1.
Figure 2. Scoring guide for Item 1.
Figure 3. Response exemplars for Item 1.
Figure 4. Evidence and Trade-offs variable map: Item threshold locations
Figure 5. Evidence and Trade-offs variable map: Criterion zones
Figure 6. Evidence and Trade-offs variable map: Comparison classrooms
Figure 7. Evidence and Trade-offs variable map: SEPUP classrooms

Many countries put chlorine in the drinking water. Chlorine is a disinfectant that is capable of killing the germs that cause diseases. However, chlorine also can react with the by-products of organic decay that are present in most water supplies to produce several suspected animal carcinogens, including chloroform.

Question #1.
Which of the following statements describes the most important issue in deciding whether or not to put chlorine in the drinking water? *Choose one.*

A. It is important to reduce the risk of getting cancer from chlorine in the drinking water.
B. It is important to prevent the spread of genes in the drinking water.
C. It is important to compare the risk of getting cancer from chlorine in the drinking water with the risk of germs in the water.
D. It is important to drink water that is not contaminated with chemicals.
E. I don't know.

Explain your answer to Question #1. You may wish to use some of the following words: *risk, cancer, chlorine, germs.*

Level 0 *Completely off topic.*

Level 1 *On topic but irrational; nonsensical; incorrect.*

Level 2 *Takes a position.*

Level 3 *Takes more than one position; cites valid supporting information.*

Level 4 *Takes more than one position; cites valid supporting information; weighs trade-offs.*

Level 5 *Takes more than one position; cites valid supporting information; weighs trade-offs; expresses uncertainty (need for more research).*

1.1 (Cholera is a dangerous disease that must be controlled.)

1.2 "Chlorine is only harmful depending on the dosage. My teacher taught as us an expression that says the dose makes the poison. So if you stay below the level of ppm then everything will be fine. But if you go over the level it can be harmful."

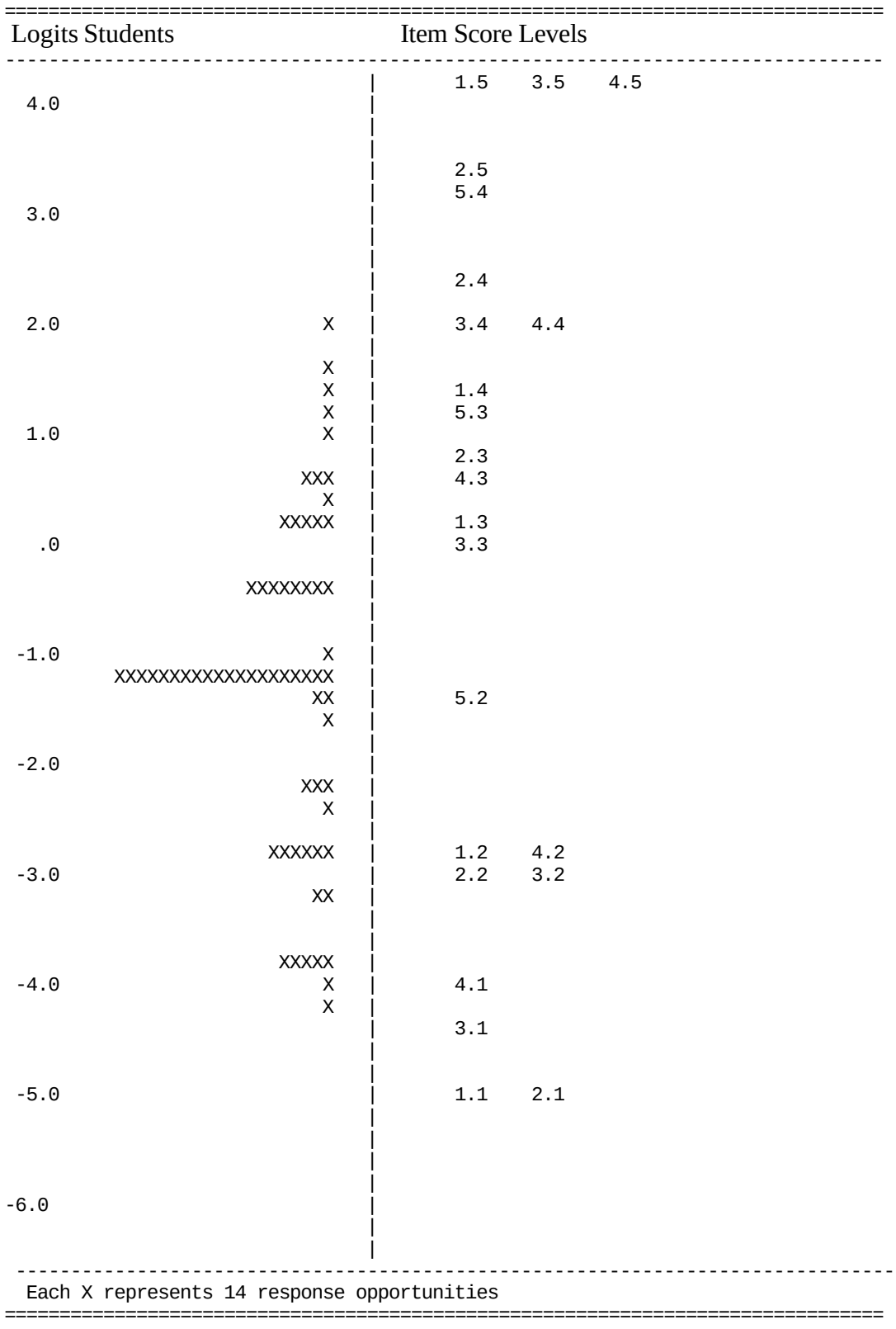
1.3 "A very small amount of Chlorine is need to kill germs (bacteria). But the risk of getting cancer is very small compared to getting disease that comes from untreated water."

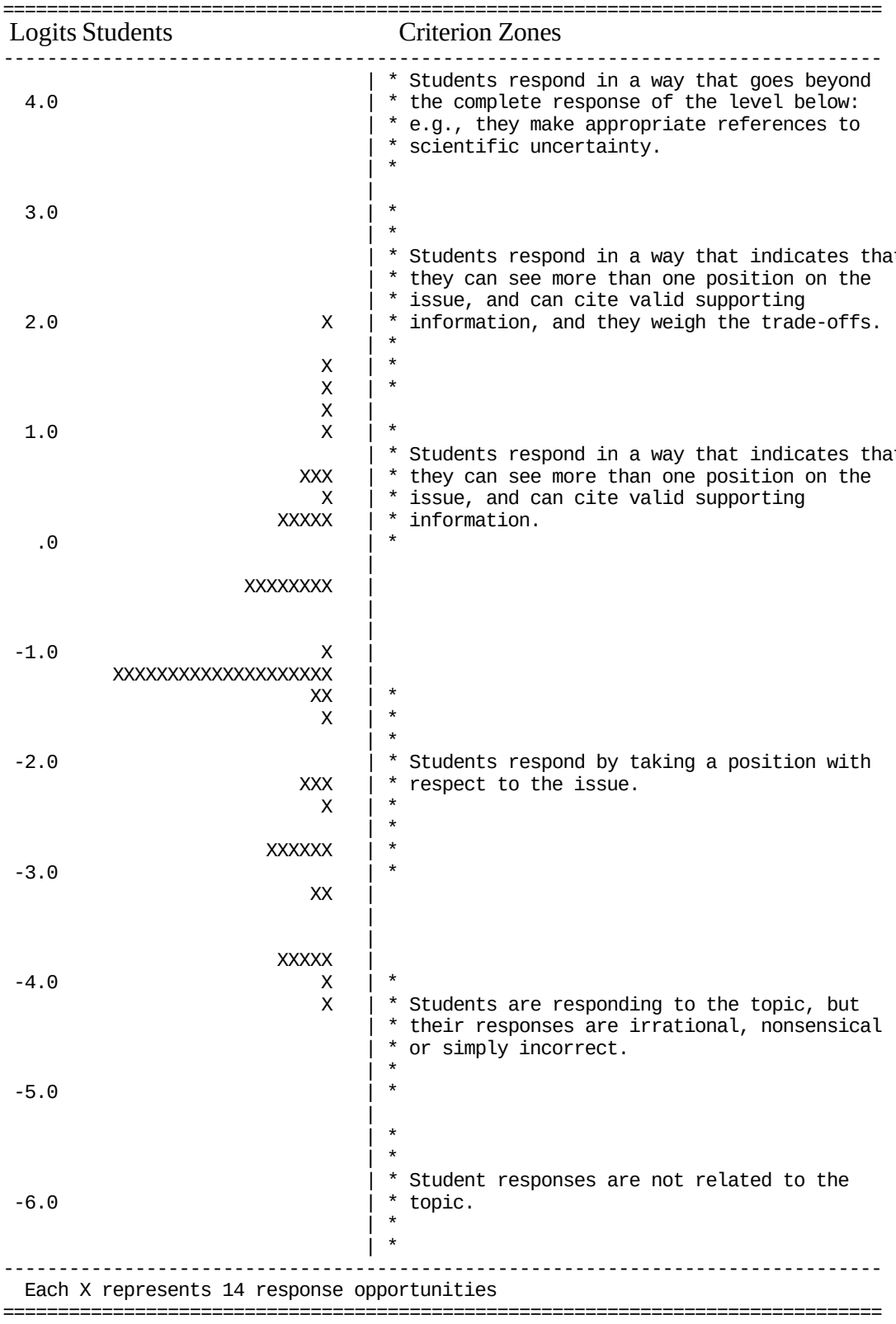
1.4 "It is hard to decide weather to put chlorine in drinking water or not. If there isn't enough chlorine, not all of the more harmful germs would be killed. If there is too much chlorine in the water, the people will get sick. There has to be an exact certain amount of chlorine in the water to deal with all of the risks involved.

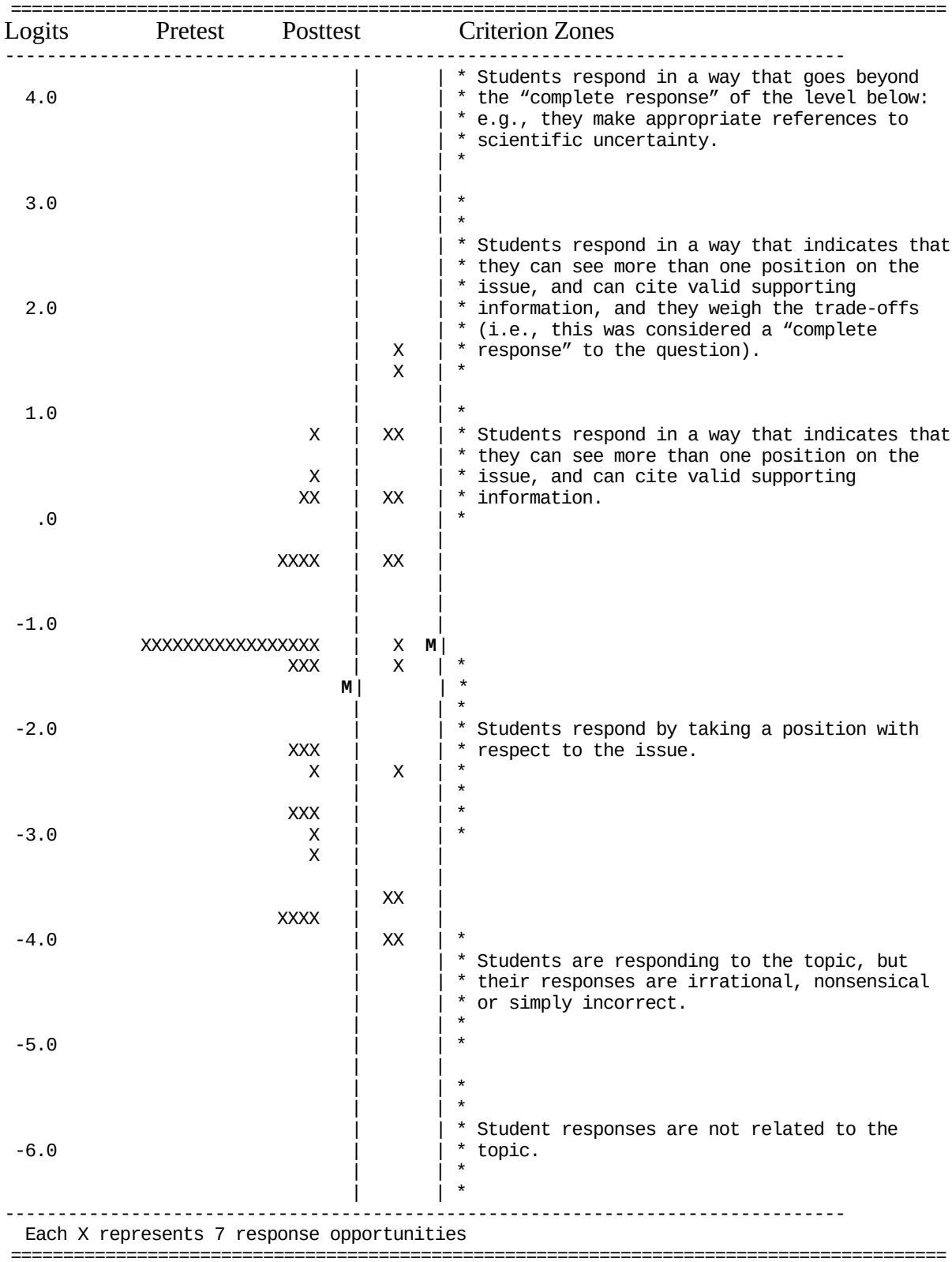
I think that companies put chlorine in water should test and make sure the amount of chlorine is acceptable, and just use their best judgment."

1.5 "Cancer is something that develops over time, but germs cause acute disease, they can harm you right away, we need to compare the risks of developing a sickness over time that does not have a known sure, yet or if we get a disease right away that can be cured, but will cost a lot of money for us if we get these germs in our system over and over again. I think that we should make a compromise on how much chlorine goes into drinking water. There should be limits. It should be tested to see how much carcinogen is produced by adding so much chlorine to drinking water and make the ultimate decision based on information of how many people would get sick, how seriously they would get sick, and how often they get sick. Again, a compromise is the answer."

(...) = exemplars in parentheses represent composites of student work--not intact quotes


Logits Students
Item Score Levels
1.5 3.5 4.5
4.0
2.5
5.4
3.0
2.4
2.0
3.4 4.4
1.4
5.3
1.0
2.3
4.3
1.3
.0
3.3
-1.0
5.2
-2.0
1.2 4.2
-3.0
2.2 3.2
4.1
-4.0
3.1
1.1 2.1
-5.0
-6.0
Each X represents 14 response opportunities


Logits Students
Criterion Zones
4.0
* Students respond in a way that goes beyond
* the complete response of the level below:
* e.g., they make appropriate references to
* scientific uncertainty.
3.0
2.0
X
* Students respond in a way that indicates that
* they can see more than one position on the
* issue, and can cite valid supporting
* information, and they weigh the trade-offs.
1.0
* Students respond in a way that indicates that
* they can see more than one position on the
* issue, and can cite valid supporting
* information.
.0
XXXXXXXX
-1.0
XXXXXXXXXXXXXXXXXXX
-2.0
* Students respond by taking a position with
* respect to the issue.
-3.0
-4.0
* Students are responding to the topic, but
* their responses are irrational, nonsensical
* or simply incorrect.
-5.0
* Student responses are not related to the
-6.0
* topic.
Each X represents 14 response opportunities


Logits
Pretest
Posttest
Criterion Zones
4.0
3.0
2.0
1.0
.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
M
M
* Students respond in a way that goes beyond the "complete response" of the level below: e.g., they make appropriate references to scientific uncertainty.
* Students respond in a way that indicates that they can see more than one position on the issue, and can cite valid supporting information, and they weigh the trade-offs (i.e., this was considered a "complete response" to the question).
* Students respond in a way that indicates that they can see more than one position on the issue, and can cite valid supporting information.
* Students respond by taking a position with respect to the issue.
* Students are responding to the topic, but their responses are irrational, nonsensical or simply incorrect.
* Student responses are not related to the topic.
Each X represents 7 response opportunities

Logits	Pretest	Posttest		Criterion Zones
				* Students respond in a way that goes beyond
4.0				* the complete response of the level below:
				* e.g., they make appropriate references to
				* scientific uncertainty.
				*
3.0				*
		X		*
				* Students respond in a way that indicates that
				* they can see more than one position on the
				* issue, and can cite valid supporting
2.0		X		* information, and they weigh the trade-offs.
				*
		XX		*
		XX		*
	X			
1.0		XX		*
	X			* Students respond in a way that indicates that
		XXX		* they can see more than one position on the
				* issue, and can cite valid supporting
	XXXX	XXX		* information.
.0				*
			M	
	XXXXXXXX	XXX		
-1.0		XX		
	XXXXXXXXXXXXXXXXXXX			
	XX			*
		M X		*
				*
-2.0				* Students respond by taking a position with
	XXXXX	X		* respect to the issue.
	X			*
				*
	XXXXXXX	X		*
-3.0				*
	XX			
		X		
	XXXX			
-4.0	X	X		*
	X			* Students are responding to the topic, but
				* their responses are irrational, nonsensical
				* or simply incorrect.
				*
-5.0				*
				*
				*
				* Student responses are not related to the
-6.0				* topic.
				*
				*

Each X represents 7 response opportunities

Table 1
Some representative values for interpreting a logit scale

Logit difference	Probability
3.0	0.95
2.0	0.88
1.0	0.73
0.0	0.50
-1.0	0.27
-2.0	0.12
-3.0	0.05

Table 2
Group means and t-tests for the SEPUP and Comparison students

Group	Pretest	Posttest	t-value	df	Prob.
SEPUP	-1.5503	-0.1462	9.39	254	.000
Comparison	-1.6314	-1.2045	1.94	124	.055