

# Assessing student learning in project-based science classrooms: Development and administration of written assessments

Phillip Herman, Scott MacKenzie, Bruce Sherin, Brian Reiser  
School of Education and Social Policy, Northwestern University  
Email: [herman@northwestern.edu](mailto:herman@northwestern.edu)

**Abstract:** This article reports on the development and administration of written pre and post assessments to help evaluate student learning in some Chicago-area classrooms using project-based science units. Written assessments administered to large numbers of students can play a meaningful role in a comprehensive evaluation effort that includes multiple methods and levels of assessment. We describe the development of assessments for two units and report preliminary results. We argue that these assessments serve multiple purposes and provide useful insights into student learning in these classrooms. Results from these assessments can help inform and focus more fine-grained qualitative analyses of student learning.

## Introduction

This article describes efforts to evaluate a reform effort for science education. In particular, we describe how we have begun to assess the learning that occurs in some middle school classrooms using project-based science (PBS) curricular units (Krajcik, Czerniak, & Berger, 1999). These units were designed by the Center for Learning Technologies in Urban Schools (LeTUS), a collaborative effort involving Northwestern University, the University of Michigan, and the public school districts of Detroit and Chicago.

Our assessment effort has faced a number of difficulties associated with evaluating modern reforms in science education. First, our reform constitutes not only a change in *how* science is taught; it is also a change in *what* is taught. The content covered is different in that it often cuts across traditional science disciplines. In addition, project-based science units emphasize the "processes" of science. Second, assessing *process* skills in an extensive, systematic manner (e.g., on written tests) poses particular challenges. Third, substantial logistical difficulties must be overcome to assess student learning. For example, over the last two years, we have attempted to gather written test data on approximately 7,200 students across 50 schools. In order to collect these assessments, we need to convince teachers to participate, have them work with us to attain signed student and parental informed consent, remind them to administer assessments before and after the units, and to return the documents to us.

This article documents our efforts to overcome these difficulties within the Northwestern-Chicago arm of our project by the development and administration of a series of written pre and posttests to assess student learning. We describe the rationale for written assessments, detail our assessment design process, present initial results from two of our units, and interpret those results within the context of a comprehensive assessment program. The two units we report on, Global Warming and Behavior Matters, were chosen because they illustrate two ends of a content-process unit design continuum. On one end of the continuum are units like Global Warming that focus heavily on students learning "content" or facts, such as "oxygen is *not* a greenhouse gas." On the other end are units like Behavior Matters that focus on scientific processes that teach students, for example, to "identify a researchable question."

## Context and history

LeTUS develops and supports the teaching of inquiry-based science units for urban middle school students in the Chicago area. Currently, LeTUS provides technical and professional development support for teachers who choose to enact any of six PBS units. LeTUS unit design is based on a model of PBS that is consistent with what we know about teaching and learning. Each unit: aligns with local, state, and national standards, is contextualized in real-world problems, allows for sustained student inquiry, embeds learning technologies, fosters collaboration, and provides educative materials for teachers (Schneider, Krajcik, & Marx, 2000).

LeTUS' early attempts to assess student learning met with limited success. Although we had successfully conducted a number of studies that carefully reported on the enactment of the units in a small number of classrooms, the kinds of adaptations teachers make when teaching LeTUS units, the ways students use technology in

investigations, and how teachers develop and apply alternative forms of assessment, we managed to collect relatively little useful data about student learning. Although in previous years we had distributed thousands of pre- and posttests, few matched pairs were returned. Systematic analysis of student learning was impossible with such an incomplete sample. Many classrooms turned in no research materials at all.

These earlier difficulties have led us to rethink our assessment efforts. First, quite simply, we needed to substantially alter our own estimates of the amount and kind of effort that is needed to get useful data. Second, we have had to be more modest in our expectations. For example, during our early attempts at assessment design we created two versions (A and B) of each assessment and asked teachers to match students from pre to post so that if a student completed the Version A pre test, the student would complete the Version B posttest. This caused considerable confusion for our teachers. We now distribute only one version of the pre and posttests that are identical at each administration. Additionally, during our early attempts at assessment, we had high hopes that we could measure deep changes in students' abilities to engage in scientific inquiry within particular domains. This led to the development of instruments that were too difficult. Students were unable to answer many of the questions on the posttest. Some open-ended items were so difficult that not even a single student attempted to answer them.

These "realizations" may seem trivial to some readers, but they represent a difficult but important shift for those in our research tradition. We believe in engaging individual students in meaningful scientific inquiry, and in fostering deep conceptual change. Within this tradition, qualitative methods focused on relatively small numbers of students are the norm. But, as we have attempted to evaluate the impact of LeTUS units on student learning in hundreds of classrooms, we have increasingly felt the need to develop and administer written tests to students, with relatively straightforward items. We believe that written tests administered to large numbers of students are necessary as one indicator of learning. They should be incorporated into a multilevel-multifaceted (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002) evaluation program that assesses student learning at various distances from unit enactment (proximal to more distal) and uses multiple methods of evaluation, fitting qualitative and quantitative results into a comprehensive model of student learning.

There were a number of factors that influenced our decision to develop written assessments of student learning. First, LeTUS stakeholders including teachers, principals, science educators, and district-level administrators have become ever more insistent on the need to document the value of participation in LeTUS. They wanted data that they could understand and use to improve teacher practice, to recruit new teachers, to train teachers, and to build parent and community support for the reform. Second, LeTUS is funded by the National Science Foundation whose representatives have stressed the need to systematically document the impact of LeTUS on student learning. Last, and perhaps most importantly, we recognized that the design and administration of these instruments would help us to better understand the enactments of the units in every classroom returning assessments. We knew too little about the enactments of "typical" LeTUS teachers. We knew something of the enactments of those maverick teachers who worked closely with LeTUS researchers, but for the majority of teachers (and particularly teachers new to LeTUS), we did not adequately understand their experiences teaching the units. The written assessments represent an entryway into their classrooms and a means of engaging teachers in a variety of our research efforts.

## **The current effort**

To date, our most systematic attempts to evaluate student learning occurred during the 2001-02 school year. Separate assessment design teams were formed for each of the 6 units being supported by LeTUS. This article reports in detail on the assessment design for two of the units, Behavior Matters (BM) and Global Warming (GW). Each team developed an assessment that reflected the particulars of each unit, as described in the next section. However, there were a number of goals and constraints that influenced the overall design of every LeTUS assessment. These included:

1. *Relatively short assessments.* Assessments were to be designed so that students could complete them in no more than one class period. This limited the number of items tested but was necessary so as not to take up too much valuable classroom time.

2. *Careful attention to the literacy demands of the assessments.* Many of our students have limited English proficiency. As much as possible, we wanted to assess students' science *achievement* and not reading *ability*.
3. *The assessments should serve multiple audiences and purposes.* Because of the investment of substantial efforts in the development, administration, collection and tracking of these assessments, we wanted them to serve as many purposes as possible. As curriculum designers and researchers, we wanted to have samples of student writing that we could refer to and study at a later time. Also recognizing that some teachers stress vocabulary in their enactments, we included multiple choice (MC) items to test vocabulary acquisition. Though the acquisition of vocabulary was not of primary interest to us, we included these items to serve these other audiences. Similarly, whenever applicable, we used released items from national standardized tests like the TIMSS as a way of interpreting the performance of our students against some national norm and also as a way of demonstrating to principals and others that the content covered in our units matches some of the content on important national assessments
4. *The assessments should be easy to score.* With limited resources, we wanted to design assessments that could be scored in a timely way so that the results could be shared with teachers and schools during the same school year. Therefore, each assessment employs multiple choice (MC) and Short Constructed Response (SCR) items scored by researchers based on a rubric.
5. *Include items of varying difficulty.* Based on our past difficulties with determining the appropriate level of difficulty for items, we made an effort to include a wide range of items that gave all students the opportunity to demonstrate what they had learned.
6. *Include items that the designers of the unit have identified as being of highest priority.* It is challenging to design assessments for scores of classrooms, particularly when it is unclear what a "typical enactment" for each unit looks like. Without understanding what teachers actually cover and what they stress, it is possible to design instruments that do not reflect students' opportunities to learn. Nonetheless, we decided that it was worth focusing on those items that the unit designers identified as being of central importance in each unit. We were therefore assured that we would be assessing learning that mattered. We return to this issue later in the paper.

### **Global Warming**

The Global Warming (GW) unit is a 6 to 12 week PBS unit designed by LeTUS in collaboration with the WorldWatcher Curriculum Project (Edelson, Gordin & Pea, 1999). The unit is designed to help students identify and understand the factors that contribute to the global warming debate by placing students in the role of advisors to different countries participating in a United Nations conference. The unit builds student understanding of important concepts in physics, chemistry, biology and geography. Students also learn to support arguments with data from the WorldWatcher data visualization software program. Students use the software and other tools to research the effects of different climate factors on a country of their choice. In the culminating activity, students present their findings and recommendations to classmates.

In comparison to other LeTUS units, GW lies closer to the "content" end of the process-content continuum. In this respect, GW is more like traditional science curricula (though it does involve substantial student inquiry). It is distinctive in the extent to which it integrates content from several scientific disciplines. Topics include: Radiative energy transfer, reflectivity, and absorption; respiration, photosynthesis, decomposition, and the carbon cycle; and Earth's energy balance, the hydrological cycle, and the greenhouse effect (Sherin, Edelson, & Brown, 2000).

Because of the complex integration of content, figuring out what "content" to assess is difficult, and assessing student learning is particularly challenging (Sherin et. al., 2000). In order to get a handle on the content in GW, we employed a framework that is being developed by Schwarz and described in another paper in these proceedings (Schwarz & Sherin). In this framework, individual components of curriculum content are described as involving one or more of a set of "epistemic forms" (Collins & Ferguson, 1993). In some cases, students learn content in the form of simple *lists*, such as the list of gases that are greenhouse gases. In other cases, students are

asked to learn and reason about *systems dynamics models*: “what would happen to the rate of global warming if the amount of clouds in the sky increased?”

The GW assessment reflects the intent of the assessment team to assess the range of content in the unit, with the particular epistemic forms in mind. For example, GW students learn about a particular *functional relationship* that holds between temperature and latitude. Students learn that temperatures are largely determined by incoming solar energy (ISE) and that the amount of ISE a location receives is a function of its position relative to the equator. Locations closer to the equator receive more ISE than those further away and therefore tend to have warmer temperatures. Using a map-like illustration, students were asked to select the location with the warmest average annual temperature and to explain their selection.

The GW assessment included 18 items. Two of the items came from released TIMMS assessments. 7 of the 13 multiple choice items (including the two TIMMS items) asked straightforward “fact”-like questions that required students to identify members of a list or to identify factual statements (“which is a greenhouse gas?” “which activity increases levels of carbon dioxide?” “fossil fuels were formed from?”, “which shape reflects sunlight?”), five (MC) and three (SCR) items asked students to explain relationships such as location/ISE mentioned above, reflectivity/groundcover, intensity of sunlight/tilt of the Earth. One (MC) item required students to interpret a table. Two (SCR) items required sophisticated reasoning about systems dynamics models as described above.

### **Behavior Matters**

The Behavior Matters (BM) curricular unit is a four to six week PBS unit designed by LeTUS in collaboration with the Brookfield Zoo (Margulis, Reiser, Dombeck, Go, Kyza, & Golan, 1999). The curriculum exposes students to methods of behavioral observation and supports them as they design and implement their own animal behavior study. The unit supports student learning through a series of lessons that help students learn to ask researchable questions and conduct background research on an animal that interests them. Students are supported in the decomposition, categorization and analysis of behaviors by the Animal Landlord software program that includes a library of digital videos and tools. Near the end of the unit, students visit a local zoo where they collect data on the animal they have selected. After the field trip, students analyze their data, discuss their findings and present these findings to their teacher and classmates. BM addresses a variety of middle school science standards, including those focused on the interaction of living things with each other and the environment, and the processes of scientific inquiry for investigating questions, conducting experiments, and problem solving.

In contrast to GW, BM lies much closer to the “process” end of the process-content continuum. In GW, there are functional relationships that can be easily quantified, and can thus be addressed in relatively straightforward assessment items. In contrast, learning to “observe behavior,” while relatively narrow in disciplinary extent, constitutes a somewhat less clear target for assessment.

One possible set of targets for assessment is particular student difficulties we have seen or that have been documented elsewhere in the literature. In describing animal behavior, middle school students often cite phenomena they have not observed or interpret actions in ways their observations do not justify. They form conclusions about animals’ emotional states that cannot be readily observed. They use simple analogies between animal and human activities; dogs and cats become capable of talking to friends, hanging out, and being selfish (Zohar and Ginossar, 1998). Reducing the tendency to anthropomorphize the behavior of animals is an important goal of the unit. Students learn that animal behavior can be scientifically observed and measured. Identifying misconceptions about animal behavior and documenting the extent to which exposure to the unit succeeds in eliminating them became a priority for the BM assessment design team.

In addition to learning why animals behave the way they do, students learn what behavioral scientists do and study the role of zoos in society. This constitutes another area about which we can ask relatively direct questions on a written assessment. The assessment asks students to identify which of four activities a scientist studying the behavior of wild condors would *not* do.

We did attempt to measure learning of the more “process”-like skills that are at the heart of BM. The unit is designed to help students learn how to identify and construct their own researchable questions. In the unit’s culminating activity, they design a habitat for an animal they have selected and present it to the class. The assessment prompts students to show how they can judge the success of a habitat designed for African wild dogs. By

the end of the unit, students are expected to favor scientific (the wild dogs in the habitat spend their time much like wild dogs in Africa do) over non-scientific criteria (The wild dogs obey the zookeeper's commands) (Item 13).

We also included items that were directed at the ability to construct scientific arguments. A premise underlying the unit is that animal behaviors reflect the costs and benefits of particular actions. It is expected that by the end of the unit students should be able to construct scientific questions and develop coherent arguments using evidence to explain behaviors they observe. An item (Item 9) designed to test this ability asks:

*Bears mostly live alone in cooler climates and in habitats with few natural predators. Why do bears live alone rather than in groups? Explain your answer.*

The BM assessment included 16 items. Four (MC) items tested vocabulary, three (SCR) items required students to interpret a data sheet that listed occurrences of animal behaviors, Four (MC) and two (SCR) items required students to reason about and explain animal behavior, one (MC) item asked students to identify animal behavior, one (MC) item asked students about what scientists studying animal behavior do. One (MC) item asked about why scientists represent data graphically.

## Scoring

Each item was awarded one point for a correct response, and zero points for an incorrect response. No partial credit was awarded. The short constructed response (SCR) items were scored using rubrics developed by the assessment design teams. These rubrics were modified once we collected a subset of student responses. One researcher then scored all of the SCR items. Two additional researchers scored a representative sample (approximately 13 %) of the items for each unit. Inter-rater reliability was assessed by averaging the pair-wise correlations among the three researchers to establish a level of confidence in the interpretability of the rubrics. Correlation coefficients ranged from .75 to .99 with a mean of .84 for Behavior Matters and .76 to .99 with a mean of .90 for Global Warming.

## Results

The two assessments were administered in classrooms in Chicago and in an adjoining suburban district that serves a diverse, but different, population of students than Chicago. We collected some demographic information about the sample but since this data was self-reported by students it is incomplete. The sample was 49.3% female, 23% Caucasian, 16% African-American, 17% Hispanic, 17% other. These breakdowns, when analyzed at the district level, roughly reflect the diversity of the student populations in the two districts. Assessments were received from 32 of the 36 classrooms enacting the GW unit and 45 of the 53 classrooms enacting BM. For GW, we collected assessments for 623 of the 1,012 students (62%) who participated in the unit. For BM, 960 of 1,335 students (72%) completed the assessments. Results are reported only for those students who turned in both pre and post assessments.

Table 1 provides summary results. Students improved their total scores by an average 2.14 points on the BM assessment and 2.04 points on the GW assessment. A dependent samples t test indicated that these differences were significant,  $p < .001$ . The effect size, .63, indicates that the average score for students on the posttest is higher than the scores of approximately 74% of the students taking the pretest. Table 1 also reports on pre and post differences by grade level. The average pretest scores for both assessments is substantially higher than what would be expected from random guessing, which indicates that students may know much of what is assessed before the unit commenced in their classrooms.

Table 2 provides mean scores for each of the items. Since the items are scored as either 0 or 1, the means on each item represent the percentage of students answering that item correctly. To determine whether the differences from pre to post were statistically significant, we used the McNemar Chi Squared test for significance of change that indicated that for 31 of the 34 items the differences were significant,  $p < .05$ . Table 2 also includes the percentage of students who incorrectly answered each item on the pretest but answered the same item correctly on the posttest.

For GW, the three largest gains from pre to post were on the first three items, each of which were MC and assessed relatively simple facts. Item 7 and Item 12 were much more challenging items because they asked students to reason about perturbations to the Earth's atmospheric system. Both items showed significant growth in student

scores, though even on the posttest only 20% of students were able to answer Item 7 correctly. Item 10 required students to interpret a table and identify a “cycle” in a dataset. Very similar activities are included in the unit, which may explain the gains on this item. For BM, the largest gains were on Items 1, 3, 4a, and 4b. Item 1 asks students to identify animal behavior, very central to the unit. Item 3, also stressed in the unit, requires students to identify the practices of scientists who study animal behavior. Items 4a and 4b are vocabulary items (“Submission” and “Forage”). Items 8 and 9, which show some change pre to post, are difficult items that require students to reason about the costs and benefits of particular animal behaviors, a concept that underlies the unit design but is not taught explicitly in specific activities. Item 12, with very little change in scores, is a very challenging item that requires students to reason about how physical characteristics and aspects of an animal’s ecosystem help to shape animal behavior.

### **Limitations**

There are a number of limitations that need to be considered when interpreting the results of our efforts. Since we argue that the LeTUS units are innovative in how and what students learn, it is difficult to make meaningful comparisons to other groups of students who did not participate in LeTUS. We have not been able to find and enlist large number of “control” classrooms that are learning similar content. Since we use the same assessments both pre and post, it is possible that students’ gains are to some degree attributable to familiarity with the tests at the posttest administration.

We have inadequate demographic information (Race, Gender, Free/Reduced Lunch Eligibility) that might better help us to understand differences in student learning as measured by these assessments. We are working to collect that data. In order to help establish the validity of our assessments, we want to compare student scores on our assessments with other measures of student achievement in science including classroom grades and scores on standardized tests of achievement. As yet, we do not have access to that data.

Item-level assessment data may be unreliable. The total scores are likely to be more reliable. We need to be cautious in how we interpret results for individual items. However, since we are using the item-level results as a guide to help us understand learning in PBS classrooms rather than as a definitive measure of absolute learning, we believe we can cautiously use item-level data in our analyses.

### **Discussion and Conclusion**

The results described here are preliminary. They are from the first of three “waves” of test results collected during the 2001-2 school year. As more tests come in for the other units, we will be better able to judge the success of our efforts. However, a few points are clear and worth noting. The statistically significant increase in student scores from pre to posttest indicated that students are learning in these units. Item-level analyses indicate that students are improving on many of the challenging items and not only on the simplest items. LeTUS has made considerable progress in collecting information about student learning in its classrooms. Compared to the return rates in prior years, the current rates represent a significant improvement. We are collecting assessments on 62% and 72% of the students enacting these units. More teachers are involved and more actively involved in our research efforts. We have provided feedback on the performance of the students on the assessments in every participating teachers classroom(s). We have thus made headway in overcoming the logistical difficulties mentioned in the introduction.

We do not claim that written assessments are sufficient indicators of learning in inquiry-based reform classrooms. To read this paper in that fashion would be a mistake. Instead, we argue that there are a variety of benefits that ensue from developing, administering, and analyzing the results of these assessments. The results of these assessments can be used to inform teacher practice, focus professional development discussions, encourage principals and other stakeholders to support the reform, and give curriculum designers feedback on specific features of the unit design. Perhaps the greatest benefit is that the process requires us to acknowledge what we know and do not know about how students and teachers make sense of and learn from these units. The development of these assessments required principled planning that does not allow for the spontaneous adjustments possible in a clinical student interview. We had to make decisions about what we valued and had to work hard to try to assess that kind of learning. The relatively high scores of students on the pretest indicate that we may have underestimated what students bring to these units in terms of prior abilities and knowledge. We are heartened that students seem to know more than we had expected. However, on certain items that we felt were central to the design of the units, students’

scores showed little or no growth. There are a number of possibilities that might explain these results. One possibility is that the items were somehow problematic and did not adequately allow students to demonstrate what they had learned. Another possibility is that students were exposed to the material but did not learn it because it was too difficult or because their teachers simply did not understand the intent of the unit. Another possibility is that the items we assessed did not reflect students' opportunities to learn. One of the best uses of the data we collected so far is as a means of gaining insight into "typical" enactments of these units by teachers. The assessments in their current form largely reflect the intent of the unit designers with some modifications and additions to serve multiple stakeholders. We now can use these results to pinpoint what we know and do not know about students' opportunities to learn and, more generally, as a way of characterizing various enactments. For example, we can look at student scores for individual teachers. If most students in those classrooms do well on the vocabulary items but not on the more "inquiry-focused" items, does that reflect what they were exposed to in that teacher's classroom? Did they complete the unit or end it only part way through? If so, that should be reflected in student scores on these assessments.

In order to better understand the enactments of these units, we are using the results of these assessments to devise new teacher surveys and interviews that will better track the enactments from the teacher's perspective. We want to eventually be able to match results on these assessments with what we know about enactments to develop a better measure of students' opportunities to learn. Without being able to judge that opportunity to learn, we are forced to make too many vague generalizations about learning in LeTUS classrooms.

Student performance on these assessments can be used to guide subsequent research that follows up on their responses. For example, if students are able to reliably choose the correct response to a MC item that asks about the causes of the seasons on Earth but cannot correctly complete the associated SCR item, we can use their responses as a way of framing subsequent student interviews that can probe more carefully on this discrepancy.

As we go forward, we will continue to assess student learning with written assessments as part of a more comprehensive model of student learning in PBS classrooms. We will work to make the assessments better and involve teachers more closely in the redesign of the assessments to ensure that the assessments reflect both the intentions of the unit designers and the reality of how they are taught in Chicago area classrooms.

**Table 1: Means and effect sizes by unit and grade level**

	<b>Global Warming</b> (Total Possible Score =18)					<b>Behavior Matters</b> (Total Possible Score =16)				
	<i>N</i>	Pre	Post	Diff	Effect Size <sup>a</sup>	<i>N</i>	Pre	Post	Diff	Effect Size
LeTUS	623	7.17 (3.07)	9.22 (3.45)	2.04*** (2.76)	.66	960	8.01 (3.41)	10.15 (3.50)	2.14*** (2.63)	.63
Grade										
05						139	7.65 (3.20)	9.54 (3.54)	1.89 (2.88)	.59
06						605	8.64 (3.44)	10.74 (3.51)	2.10 (2.56)	.61
07	31	4.90 (2.18)	7.45 (2.55)	2.55 (3.00)	1.17	141	6.01 (2.66)	8.70 (2.89)	2.68 (2.44)	1.0
08	592	7.29 (3.07)	9.31 (3.47)	2.02 (2.74)	.66	75	7.35 (3.20)	9.25 (3.30)	1.91 (2.96)	.60
<sup>a</sup> Effect size: Effect size was calculated by the difference between the means divided by the pretest standard deviation.										
*** p<. 001.										

Table 2: Item level results

Global Warming					Behavior Matters				
Item	Type <sup>a</sup>	Means <sup>b</sup>		% Increase pre to post <sup>c</sup>	Item	Type	Means		% Increase pre to post
		Pre	Post				Pre	Post	
01	MC	.14	.56***	42	01	MC	.33	.56***	24
02	MC	.56	.73***	18	02	MC	.59	.70***	11
03	MC	.27	.53***	27	03	MC	.48	.67***	19
04	MC	.35	.42***	07	04a	MC	.41	.66***	24
05a	MC	.60	.64	03	04b	MC	.43	.78***	35
05b	SCR	.30	.39***	09	04c	MC	.66	.80***	14
06	MC	.71	.78*	07	04d	MC	.79	.90***	11
07	SCR	.09	.20***	12	05	MC	.46	.57***	11
08a	MC	.76	.86***	10	06	MC	.48	.56***	08
08b	SCR	.62	.76***	15	07	MC	.63	.65	02
09a	MC	.26	.26	0	08	MC	.42	.52***	10
09b	SCR	.15	.18*	03	09	SCR	.40	.55***	15
10	MC	.50	.65***	15	10	SCR	.77	.85***	09
11	MC	.26	.32**	07	11	SCR	.32	.43***	11
12	SCR	.20	.33***	13	12	SCR	.13	.19***	06
13	MC	.66	.73*	07	13	MC	.60	.76***	15
14	MC	.53	.66***	13					
15	MC	.16	.22**	06					

<sup>a</sup>Type: MC = multiple choice, SCR = short constructed response.

<sup>b</sup>Means: Items are scored as 0=Incorrect or 1=Correct so the means represent percentage of students correctly answering each item.

<sup>c</sup>Percent Increase: Percent of students who answered pre item incorrectly but answered correctly on the posttest.

\*p<. 05; \*\* p<. 01; \*\*\* p<. 001.

## References

- Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, 28(1), 25-42.
- Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the Challenges of Inquiry-Based Learning Through Technology and Curriculum Design. *Journal of the Learning Sciences*, 8(3&4), 391-450.
- Krajcik, J., Czerniak, C., & Berger, C. (1999). *Teaching Children Science: A Project-Based Approach*. McGraw-Hill College Press.
- Margulis, S. W., Reiser, B. J., Dombeck, R., Go, V., Kyza, E. A., & Golan, R. (2001). Behavior Matters: Involving Students in Scientific Investigations of Animal Behavior. Presented at the 2001 Annual Meeting of the national Association for Research on Science Teaching, St. Louis, MO.
- Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the Evaluation of Systemic Science Education Reform: Searching for Instructional Sensitivity. *Journal of Research in Science Teaching*, 39(5), 369-393.
- Schneider, R. M., Krajcik, J., & Marx, R. W. (2000). The Role of Educative Curriculum Materials in Reforming Science Education. In B. J. Fishman & S. F. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences*. Mahwah, NJ: Erlbaum.
- Sherin, B., Edelson, D. C., & Brown, M. (2000). Learning in task-structured curricula. In B. J. Fishman & S. F. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences*. Mahwah, NJ: Erlbaum.
- Zohar, A., & Ginossar, S. (1998). Lifting the taboo regarding teleology and anthropomorphism in biology education—Heretical suggestions. *Science Education*, 82, 679-697.