Maryland Virtual High School CoreModels Project
Student Assessment Results
Internal Report

Discussion

For the past three years, the Maryland Virtual High School (MVHS) CoreModels project has collaborated as a community of high school science teachers throughout Maryland to develop and implement instructional activities built around computer modeling in STELLA, a modeling language with an icon-based interface. Each activity was designed to meet the Maryland High School Science Core Learning Goals as well as the AAAS Project 2061 Benchmarks. Teachers reported that, as a result of using the materials, they saw improvement in their students' ability to:

- meaningfully interpret the graphical representation of data
- understand the ability of a model to represent real world behavior

At the same time that MVHS teachers were implementing these modeling activities, the Maryland State Board of Education was field-testing the Maryland High School Assessment (HSA) tests, the final piece of the state's systemic reform plan. The HSA includes both selected response items (i.e. multiple-choice) and constructed response items which require the analysis, synthesis, and written expression of ideas. MVHS teachers were concerned that, because their students were not accustomed to expressing themselves in writing in science classes, they would have difficulty with the constructed response items. Therefore, MVHS decided to assist the teachers in providing practice to their students by using the constructed response mode to measure student understanding gained through modeling.

Methods:

As an initial study, we sought to determine whether the teacher observations listed above were actually measurable. Two open-ended questions were designed for each activity in biology and physics. The first question presented the student with a graph from the topic recently studied and asked the student to explain its meaning. The second question asked the student to evaluate the ability of the model to represent real world behavior. Both questions would be scored using the 5-point Maryland High School Science Rubric, the same one to be used on the High School Assessment exams. Since the teachers reported seeing improvement in their students' ability to meet these learning objectives as more MVHS activities were used in classroom instruction, we decided to measure the effect of multiple uses of MVHS activities on student learning.

In the fall of 1999, we asked for teachers who could meet the following conditions:

1. Cover three MVHS activities during the second semester and administer an assessment after each one.
2. Send the original assessments to MVHS and keep a copy to return to their students.
3. Score the copies according to the Maryland High School Science rubric.

4. Return the scored copies to the students and discuss the answers before doing the next assessment.

Eleven biology teachers and four physics teachers responded to our request. The teachers' experience with STELLA ranged from 1 to 4 years, and the classes ranged from Basic Skills to Advanced Placement. Eleven schools were represented in the study, 6 rural, 4 suburban , and 1 urban. In the summer of 2000, these teachers met together to do a formal scoring of the assessments. Each question was subjected to blind scoring by two teachers, with a third teacher resolving discrepancies. (We need inter-rater reliability here.)

Results - Biology

The time that passed between consecutive assessments ranged from same day to 12 weeks. The assessments covered 8 different topics, of varying levels of content difficulty. Teachers did not teach the topics in the same order. Since the sequence of quiz topic administration was counterbalanced across all classes, conceptual difficulty was not a confounding factor in the analysis done to assess the effect of exposure to modeling activities on students' performance.

Although the biology teachers tried to meet the goal of three assessments, four of the eleven teachers only gave two assessments. In Charts 1 and 2, the plots in blue show the means for question 1 and question 2 for the 358 students who took at least two assessments. The plots in magenta show the means for the 160 students who completed three assessments. The plots in green show the means for one particular teacher who will be described later.

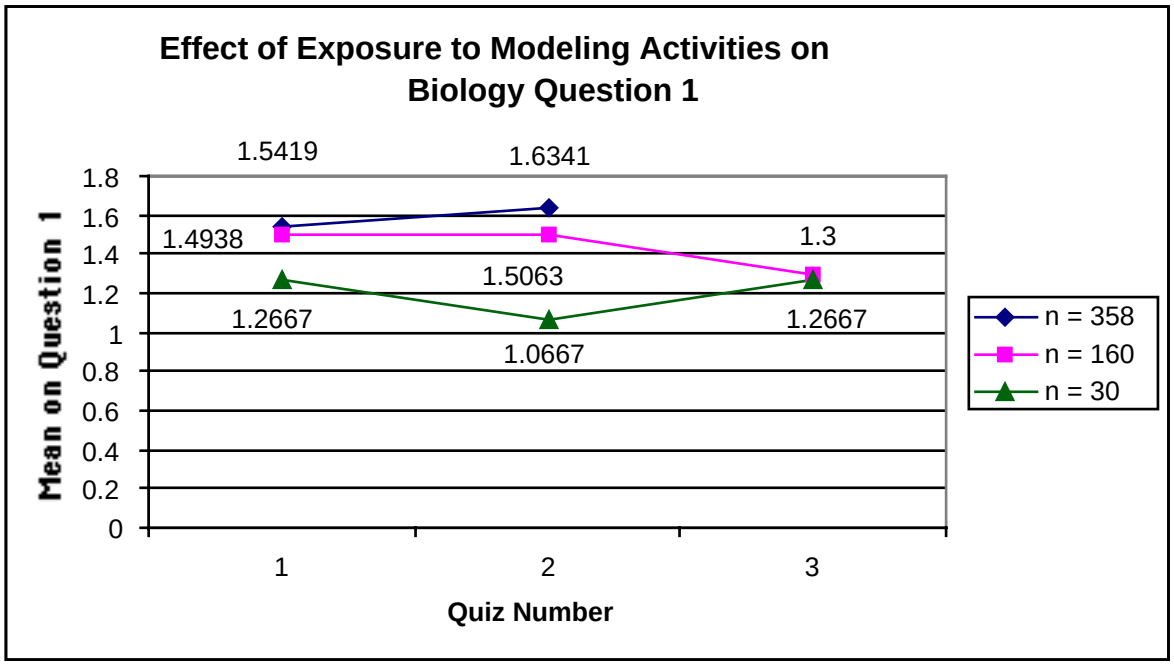

Chart 1

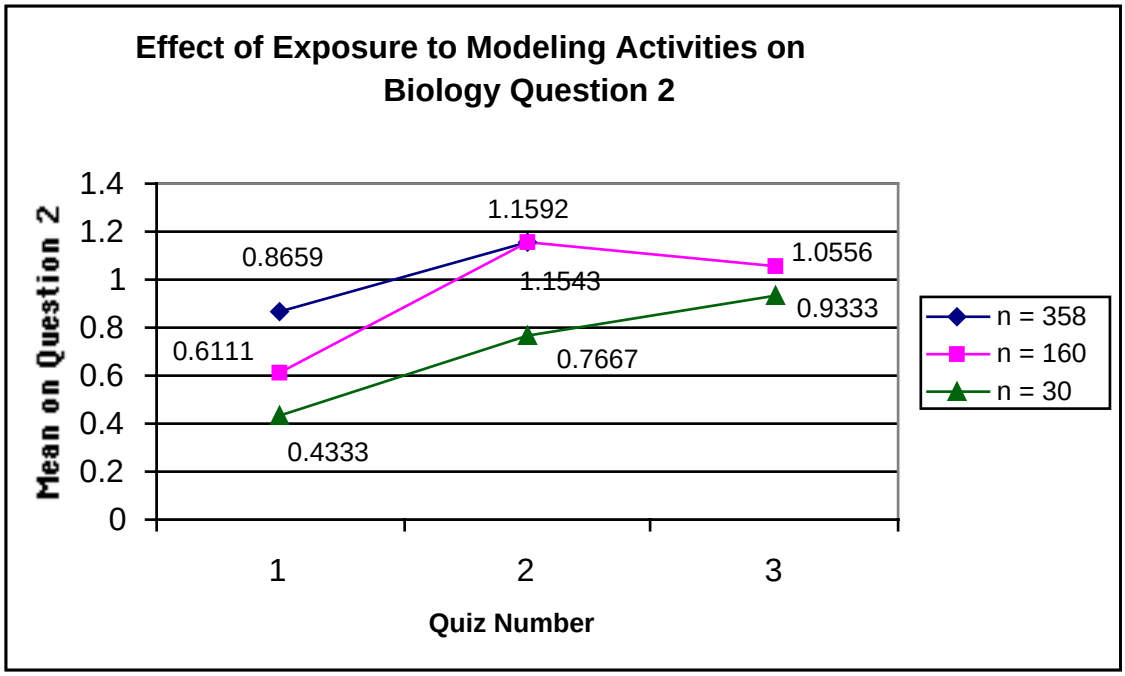

Chart 2

The tables below list the means that are described in the following analysis.

STUDENT MEANS FOR QUESTION 1

	N=358	N=160	N=30
QUIZ 1	1.5419	1.4938	1.2667
QUIZ 2	1.6341	1.5063	1.0667
QUIZ 3	NA	1.3000	1.2667

STUDENT MEANS FOR QUESTION 2

	N=358	N=160	N=30
QUIZ 1	0.8659	0.6111	0.4333
QUIZ 2	1.1592	1.1543	0.7667
QUIZ 3	NA	1.0556	0.9333

The Repeated Measures ANOVA Pillai's Trace test was used to measure the effect of exposure to modeling activities on student scores. For n=358, as exposure to modeling activities increased, students' performance on question 1 of the quizzes administered increased significantly ($p = .055$). Students' performance on question 2 also increased significantly. ($p = .000$).

Although these results appear to support the hypothesis that student achievement increases with repeated exposure to modeling activities, analysis of the data from the students who completed three quizzes (n=160) does not support such a conclusion. The means for question 1 indicated that with more exposure to modeling activities, students' performance dropped significantly ($p = 0.005$). The means for question 2 indicated that with more exposure, students' performance increased significantly from quiz 1 to quiz 3 ($p = 0.000$), although there was a small drop from quiz 2 to quiz 3. These results raise the question: Why do the scores take a downward turn on quiz 3 instead of maintaining their upward trend?

One possible explanation is that most teachers gave their third quiz after mid-May, a time when students are less likely to try their best because it is so close to the end of the school year. However, one teacher gave the third quiz in April. To test the hypothesis that time of school year had an effect on student performance, the 30 students who took their third quiz in April were analyzed (the plot in green). The mean for question 1 exhibited a different pattern from the others, going down after the first quiz and up again for the third. The mean for question 2, however, exhibited a significant upward trend. It is interesting to note that the 30 students represented two different classes and two different levels, one class of basic skills and one on-level. In both classes, the second quiz covered a topic which involved more abstract concepts than did the topic covered by the first quiz.

Discussion of Results - Biology

Question 1 involved graph interpretation skills. Although the mean increased a small amount initially, it is not clear whether the increase was due to increased exposure to modeling activities or the influence of already having a similar question on the previous quiz. Of course, there is also the influence of other student learning occurring in math and science. Question 2 required a written description of ways in which the computer model was similar to and different from the phenomenon it was meant to represent. The results there are more promising. Even when the majority of third quizzes were given late in the school year, the mean for quiz 3 was higher than the mean for quiz 1. However, it is still unclear whether the effects were due to increased exposure to modeling activities or some other factor.

We wonder if teacher comfort with a question type has any correlation with student performance on that question type. For example, graph interpretation is an area in which many biology teachers have difficulty using mathematically accurate terminology themselves. The teachers recognize this weakness in their backgrounds and are eager for more opportunities to practice graph interpretation skills with their students. Although the modeling activities provide that practice, it is possible that teacher reinforcement in classroom discussions needs to be improved in order to see steady improvement in student performance. We cannot expect to see student gains if their teachers are not clear in their own expression of the meaning of graphs.

On the other hand, question 2 simply requires a written description of the similarities and differences between the model and the real world. We know from anecdotal evidence from the classroom that teachers do increase their focus on model interpretation skills after the administration of the first quiz. Therefore, the large increase in mean scores from quiz 1 to quiz 2 is at least partially attributable to teacher focus and student familiarity with the type of question and the way in which it is scored. The challenge that remains is the identification of the factors which influenced the mean score for quiz 3.

Results - Physics

The time that passed between assessments ranged from one week to 13 weeks. The assessments covered 6 different topics, of varying levels of content difficulty. Due to the sequential nature of the physics curriculum, everyone covered the same topic for Quiz 1 (Simple Kinematics). However, the four teachers then varied in which of the other models and associated quizzes they chose to use.

Three of the four teachers had classes in schools using a block schedule in which each class meets for a 90 minute period each day, covering a year's worth of physics in one semester. Since the assessments were being administered during the second semester of the year and the first topic of the physics course is simple kinematics, the MVHS activities fit seamlessly into the course content. The fourth teacher volunteered to use the assessments also even though his classes met for a single period a day and had started in the fall. He decided to use the MVHS activities as review materials for topics already covered. Since the usage of the materials for the fourth teacher was different from the others, the data have been analyzed separately.

In Charts 3 and 4, the plots in blue show the means for question 1 and question 2 for 91 students representing 4 classes taught by three teachers at three different schools. The other plots are for the teachers individually and will be discussed later. The Repeated Measures ANOVA Pillai's Trace test was used to measure the effect of exposure to modeling activities on student scores. For n=91, as exposure increased, students' performance on question 1 of the quizzes administered decreased significantly ($p = .026$). On the other hand, students' performance on question 2 increased significantly with increased exposure to modeling activities ($p = .005$).

As exposure to modeling activities increased, it was expected that student achievement would increase on both quiz questions. The data did not support this hypothesis. We observed that the means for the graph interpretation question went down, while the means for the model interpretation question went up. To explore possible reasons for this discrepancy, we looked at the content of the quizzes and discovered three different teacher behaviors. We then found the student means for those three teachers.

The plots in magenta (n=44) represent a teacher whose quizzes covered simple kinematics, drag, and universal gravitation - each quiz introduced a new concept. The plots in green (n=22) represent a teacher whose quizzes covered simple kinematics, dynamics, and drag - the second quiz introduced the concept of force, and the third quiz reinforced that concept. The plots in red (n = 25) represent a teacher who gave eight quizzes in all, the first three all reinforcing kinematics - simple kinematics, free fall, and braking distance.

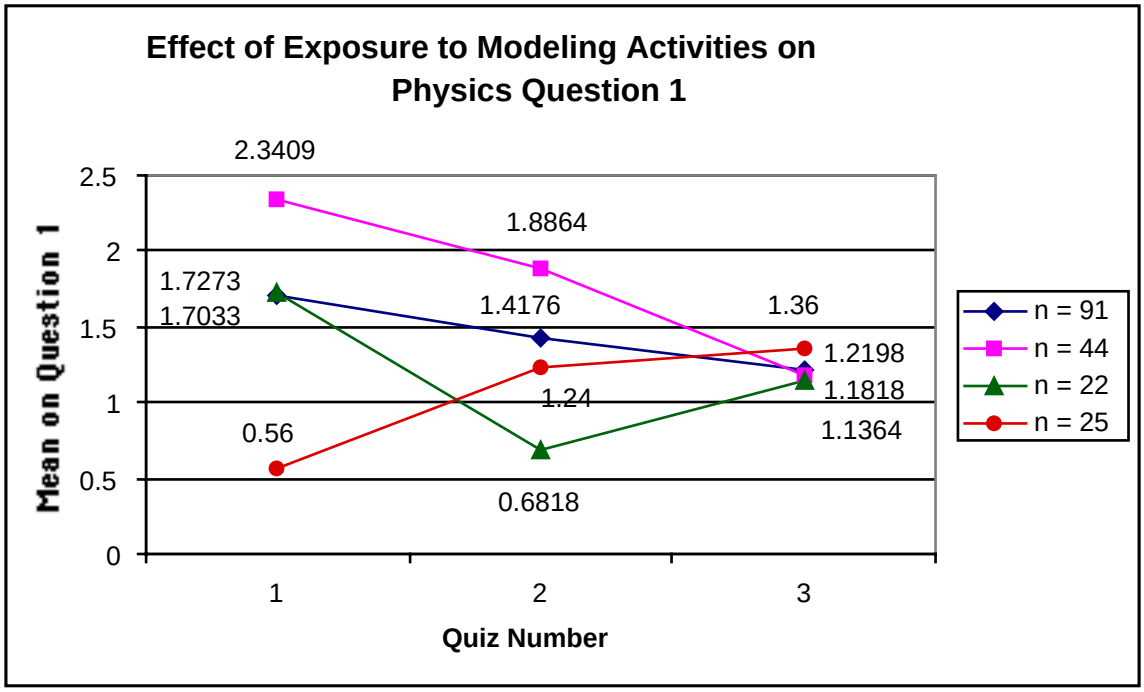

Chart 3

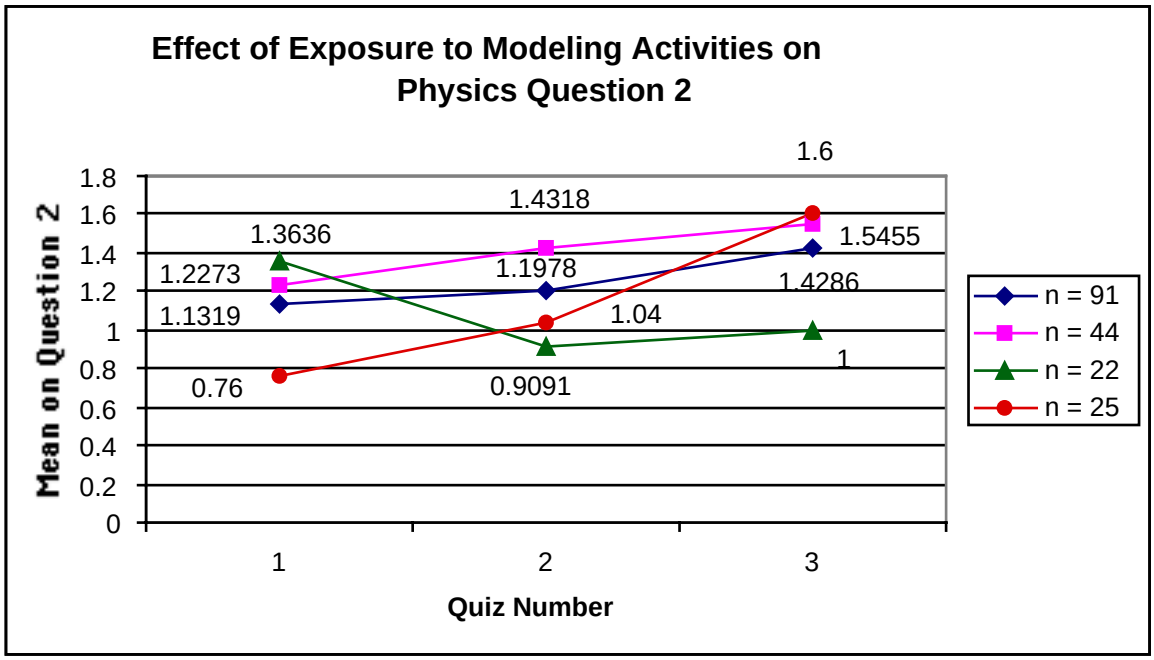

Chart 4

Why do the student means on question 1 vary so from teacher to teacher? Why do the student means on question 2 behave somewhat differently? It would appear that students' graph interpretation skills do not correlate with their ability to explain the relevance of the model to the real world. Does the introduction of a new concept have a greater impact on graph interpretation skills than on model interpretation ability? What are the important cognitive issues at work here? What other differences need to be accounted for?

The teacher who gave eight assessments (n=24) gives us the opportunity to look at the interplay between exposure to modeling activities and the difficulty of the specific topic being covered. This teacher saw student performance increase significantly on both questions 1 and 2 over the first four quizzes covering kinematics related topics. When the concept of force was introduced (quiz 5), the student means dropped dramatically on question 1, but less so on question 2. Elevator, the topic covered on quiz 6, reinforced the concept of force. The student means on both questions 1 and 2 increased significantly. Therefore, it seems likely that the introduction of a new concept may play an important role in student assessments in spite of the number of previous exposures to modeling activities. Quiz 7 and quiz 8 were given near the end of the school year and covered two very different topics, so it is difficult to determine the reasons for the student scores on those assessments. Charts 5 and 6 show the plots of the means for this teacher's class.

STUDENT MEANS

QUIZ: TOPIC	QUESTION 1	QUESTION 2
QUIZ 1: Simple Kinematics	0.5833	0.7917
QUIZ 2: Free Fall	1.2083	1.0417
QUIZ 3: Braking Distance	1.3333	1.6250
QUIZ 4: Tailgating	1.4583	1.5833
QUIZ 5: Dynamics	0.6250	1.3750
QUIZ 6: Elevator	1.1250	2.1667
QUIZ 7: Drag	1.0417	1.2917
QUIZ 8: Universal Gravitation	0.6250	1.2917

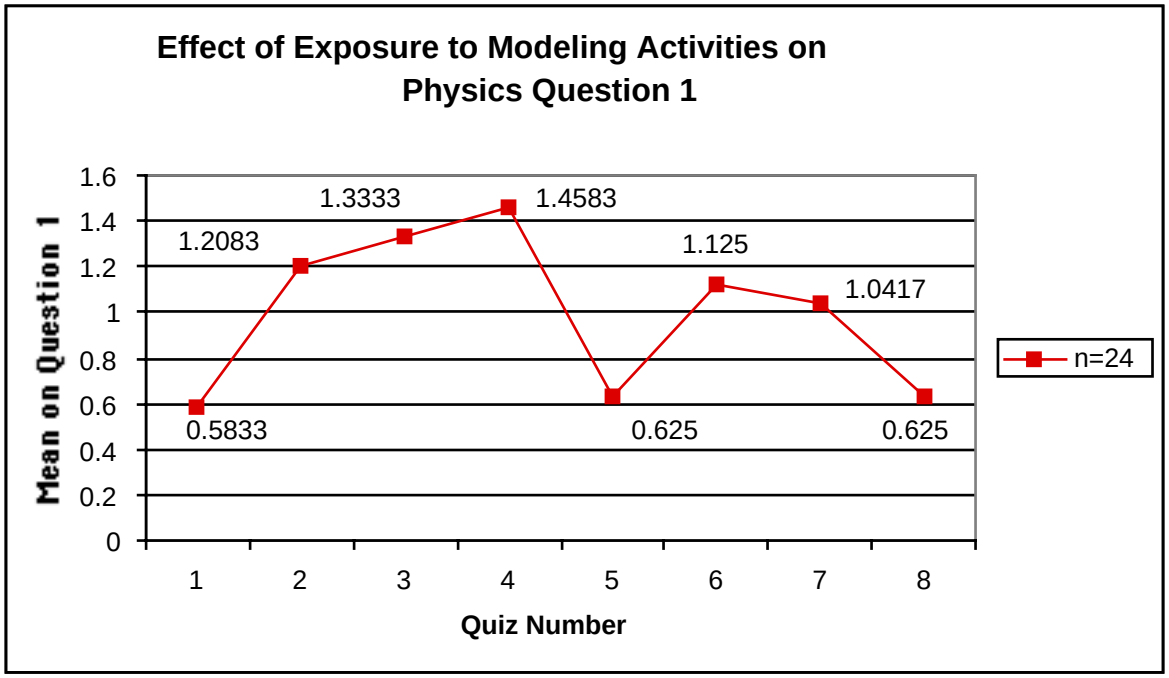

Chart 5

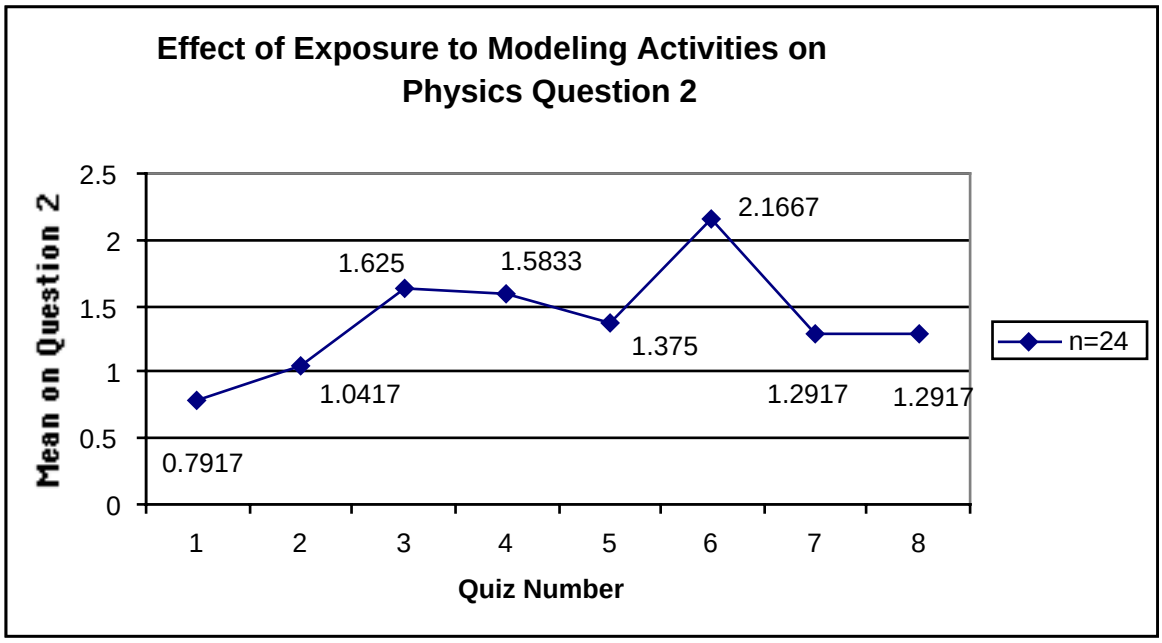

Chart 6

The fourth teacher, who used the MVHS activities to review topics already covered earlier in the year, saw similar results to the others. The first two quizzes covered kinematics topics - simple kinematics and freefall. As exposure to modeling activities increased, the performance of the 71 on-level and honors students increased on question 1 from a mean of 1.8310 on quiz 1 to 1.9436 on quiz 2. The mean on question 2 increased from 1.2113 on quiz 1 to 1.5775 on quiz 2. The third quiz, dynamics, not only covered the concept of force, it was also given nearly the last day of school. The mean on question 1 for quiz 3 was 1.0845 and the mean for question 2 was 0.4507.

Discussion of Results - Physics

In looking over the physics results, it is apparent that more data is needed in order to formulate any conclusions. Question 1, which involved graph interpretation skills, appears to be more highly sensitive to the effect of content than question 2. Although the increase in content difficulty has some explanatory value in the decrease in scores in moving from kinematics to dynamics, other factors cannot be dismissed. Question 2 required a written description of ways in which the computer model was similar to and different from the phenomenon it was meant to represent. The results there are more promising. However, the interaction of repeated quizzes with repeated modeling activities must be investigated before claiming any causality for exposure to modeling activities.

Future Directions

Many questions about the measurement of student learning have been raised by these results. It is likely that we need a multi-pronged approach to capture the true picture.

1. Should we try to identify more specifically the cognitive skills that the modeling activities directly address in order to answer the question - what kinds of learning do modeling activities enhance? Should we use a rubric that allows finer granularity?

2. Should we videotape student discussions in order to gain an understanding of the student thinking provoked by the models?

3. Should we ask more questions on our assessments? Should there be content questions, multiple-choice questions, and open-ended questions? Should we be asking all of the kinds of questions that a teacher would typically ask on a test rather than restricting ourselves to two open-ended questions?

4. Should we run an experiment with control groups in which we ask the same kinds of questions with and without the intervention of modeling activities? Should we include pre and post-tests in the experiment?

5. Should we try to compare classes based on end-of-semester or end-of-year exams? We would have to identify the questions that should benefit from modeling activities as well as those questions that are unrelated to modeling. For each set of questions, the mean of the control classes would be compared to the mean of the treatment classes. Control and

treatment classes would be comparable in course levels and teacher experience. If the treatment classes scored as well as the control classes on questions unrelated to modeling and scored better than the control classes on questions related to modeling, the treatment would be judged successful.