The impact of the MARS curriculum on students’ ability to coordinate theory and evidence

Corinne Zimmerman, Department of Psychology, Illinois State University, Campus Box 4620, Normal, IL 61790–4620, USA; e-mail: czimmer@ilstu.edu, Kalyani Raghavan and Mary L. Sartoris, Learning Research and Development Center, University of Pittsburgh, USA

The Model-Assisted Reasoning in Science (MARS) project seeks to promote model-centered instruction as a means of improving middle-school science education. As part of the evaluation of the sixth-grade curriculum, performance of MARS and non-MARS students was compared on a curriculum-neutral task. Fourteen students participated in structured interviews in which they experimented with a balance apparatus that provided three manipulable variables (two affected balance, one was a non-causal distractor variable). Although both groups were equally able to identify and test variables, all MARS students discovered a quantitative rule to describe the operation of the balance, whereas only one non-MARS student did so. MARS students discovered this numerical relationship through experimentation, regardless of their scientific reasoning profile (i.e. theory-generating, theory-modifying, or theory-preserving). The critical components of MARS instruction that may foster the ability to flexibly coordinate theory and evidence include multiple opportunities to draw conclusions from data and an emphasis on the successive refinement of models.

Introduction

Since 1991, the Model-Assisted Reasoning in Science (MARS) project at the University of Pittsburgh’s Learning Research and Development Center has worked to investigate and promote model-centered instruction as a means of improving middle school science education (Raghavan, Kesidou and Sartoris 1993, Raghavan and Glaser 1994, 1995, Raghavan, Glaser and Sartoris 1995). In MARS instruction, students learn to use models the way scientists use them, as communication and reasoning tools that can help them describe physical-world phenomena, to depict and test ideas about underlying causes, and to identify and explore relationships between ideas. In this paper, we report research designed to evaluate the impact of the MARS program on sixth-grade students’ scientific inquiry skills.

Previous research examining the MARS curriculum focused on documenting students’ prior ideas and emerging conceptual understanding (for example, Raghavan and Glaser 1995, Raghavan, Sartoris and Glaser 1998a, 1998b), with the goal of refining instruction and developing the curriculum. Current research is focused on evaluating the MARS program relative to established science curricula to determine in what ways this program fosters different aspects of student thinking in science. The evaluation included a large-scale, multi-site field test that involved the administration of paper-and-pencil tests of scientific reasoning and concept understanding to over 700 students (Raghavan and Sartoris 2003, Raghavan,
Sartoris and Zimmerman 2002), and blind performance interviews conducted with a small group of MARS and non-MARS students as they completed a hands-on inquiry task. The interview data are the focus of the present paper. We begin with a brief description of the impetus for and development of the MARS curriculum, and summarize the current instructional module for the sixth grade.

Rationale for the development of MARS instruction

The MARS program was created as a proposed instructional solution to recent trends at the middle-school level (Grades 6–8, ages 12–14). In the Third International Math and Science Study (TIMSS), the US ranked near the top in science achievement among third and fourth graders (Martin, Mullis, Beaton, Gonzalez, Smith and Kelly, 1997). In contrast, between the fourth and eighth grades, positive attitudes toward science decline and negative attitudes increase, especially among female and minority students (Bae, Choy, Geddes, Sable and Snyder 2000, Beaton et al. 1996, Mullis and Jenkins 1999, Weinbergh 1995). Although science achievement scores of 13 year olds on the National Assessment of Educational Progress (NAEP) have risen somewhat since 1977, improvement has been slow and sporadic (Campbell, Hombo and Mazzeo 2000, Campbell, Reese, O’Sullivan and Dossey 1996, NAEP 1992, 1994). TIMSS 1999 results provide further evidence that science achievement declines from elementary to middle school. The population of students who performed so well as fourth graders in 1995 did not do nearly as well when tested as eighth graders in 1999. A preliminary report on NAEP’s 2000 science assessment shows no improvement in the racial/ethnic performance gap and shows a widening of the gap favoring males over females (NAEP 2001).

At the middle school level, instruction emphasizes content more than process, and the content is often abstract and complicated. There are fewer opportunities for hands-on experience and, when such activities are used, the focus is not on visible traits and processes, but on underlying principles that are not directly observable. The are many possible reasons why enthusiasm for science at the middle school level wanes and performance falters, but one candidate could be that this transition from concrete to abstract is too abrupt. Perhaps students who once loved hands-on tasks in elementary school may suddenly find them uninteresting because they cannot see the underlying concepts. Young adolescents often have difficulty reasoning with abstract ideas, particularly when they must distinguish between observable events and unobservable properties or interactions that cause them (Bliss 1995, National Research Council [NRC] 1996). One proposed solution is instruction that fosters qualitative understanding by providing a cognitive bridge between the concrete and the abstract.

The MARS curriculum: sixth grade module

The MARS science curriculum uses models to provide this bridge. A model is an object, drawing, diagram, or some other means of representing an idea, object, process, or system (Raghavan and Glaser 1995). Models can make abstract entities visible and concrete, and they can simplify complex phenomena, omitting all but the essential features or components. Models can depict relationships, organizing key elements of a system into a coherent structure. Each component can thus be viewed...
not as an isolated fragment, but as part of an integrated system, with clearly defined connections to other elements within that system. By focusing on representations that simplify, concretize, and integrate abstract ideas, model-centered instruction can help students understand science concepts (Wiser, Kipman and Halkiadakis 1988, Mayer 1989, Smith, Maclin, Grosslight and Davis 1997, Frederiksen and White 2000, Treagust, Chittleborough and Mamiala 2002).

In addition to learning concepts, students must understand the processes by which scientific ideas are generated. In the National Science Education Standards, models are included among the unifying concepts and processes that offer ‘insightful ways of thinking about and integrating a range of basic ideas that explain the natural and designed world’ (NRC 1996: 115). Scientists routinely use models to depict and test theories, and science education researchers advocate that students gain experiential insight into this process by using models to depict and test their own ideas (Bruer 1993, Clement 1989, Glaser 1984, Hesse 1966, Hestenes 1987, Justi and Gilbert 2002, Kaput 1991, Lesh and Lammon 1992).

The core instructional component of MARS is a set of computer tasks in which students work with visible models of abstract concepts. Hands-on and written activities augment the software by embodying a concept or a set of concepts that students explore at the computer. In essence, the computer models provide ‘hands-on’ experience with abstract concepts, enabling students to manipulate symbolic representations and receive appropriate feedback. The models help students make sense of hands-on activities, providing a bridge between experience and theory. Moreover, by enabling students to explore relationships between concepts, the models foster construction of a coherent explanatory system. By learning to use models as explanatory and reasoning tools, students build qualitative understanding of science concepts and, equally important, develop scientific modes of thought and analysis (Raghavan, Kesidou and Sartoris 1993, Raghavan and Glaser 1994, 1995, Raghavan et al. 1995).

Topics covered in the sixth-grade module of the MARS curriculum include states and properties of matter, area, volume, mass, density, interaction properties, force, net force, and equilibrium. These topics correspond to concepts recommended for middle-school students in Benchmarks and National Science Education Standards (American Association for the Advancement of Science 1993, NRC 1996). They were selected for the MARS curriculum because they are well suited to the design of a coherent instructional system that supports the development of model-assisted reasoning. Researchers have identified the intuitive ideas individuals have about many of these topics, enabling curriculum developers to challenge misconceptions and build on preconceptions, and to create instruction that is intelligible, relevant, and leads to an appreciation of scientific theory building as an incremental and corrective process (Driver, Guesne and Tiberghien 1985, Duit, Goldberg and Niedderer 1992, Solomon 1992, Camp and Clement 1994, de Vos and Verdonk 1996, NRC 2000, Jones, Carter and Rua 2000).

The MARS sixth-grade module provides a gradual introduction to model-assisted reasoning, a sophisticated and multifaceted process skill that requires progressive learning, explicit instruction and extensive practice (Harrison and Treagust 2000, Lehrer and Schauble 2000, Justi and Gilbert 2002). Students learn to work with concrete, analogical models such as area tiles and volume cubes before encountering more abstract, scientific models including density dots and force diagrams. Model manipulation likewise progresses throughout the module.
Students begin by physically moving objects and computer images. Gradually, they learn to perform visual and symbolic manipulations, generating and using drawings, diagrams, formulas and example objects to represent and reason about science concepts. MARS instruction repeatedly engages students in such activities as comparing and evaluating different models of the same thing, identifying similarities and differences between a given model and the thing it represents, recognizing patterns in data, drawing conclusions from data, and using data as evidence for or against a given statement. In the final sixth-grade unit, students conduct open-ended experiments in which they use models to depict and test their own ideas about the forces that act when objects are hung from springs.

**Evaluating the MARS curriculum**

Based on information gathered from previous studies (for example, Raghavan et al. 1998a, 1998b), the curriculum was refined with some new topics included (e.g. states and properties of matter) and such topics as buoyant force and flotation moved to the seventh-grade module. The revised sixth-grade module underwent multi-site field testing in the 1999–2000 and 2000–01 school years. Evaluation instruments were administered to non-MARS students to provide normative data (Raghavan and Sartoris 2003, Raghavan et al. 2002). Non-MARS sites had been involved in a local systemic change project, implementing nationally recognized hands-on science modules for 4 years.

As mentioned previously, field tests included two forms of a paper-and-pencil test of scientific reasoning developed by Lawson (1987) that measure specific reasoning skills, including conservation, proportional thinking, identification and control of variables, and probability. The second component was a written concept test made up of released items from TIMSS and other published instruments. For these measures MARS students improved significantly from pre-instruction to post-instruction, and for many measures the MARS students outperformed the non-MARS students (for details, see Raghavan et al. 2002).

In addition to these large-scale quantitative studies, the evaluation included blind interviews conducted with a small group of MARS and non-MARS students from one of the implementing schools. Students were interviewed while they experimented with a balance to determine which variables did and did not affect the operation of the balance. This smaller, qualitative study was used to gain insight into the ways that students reason online during a hands-on inquiry task to supplement the findings of the written instruments.

Although MARS is a science curriculum focused on conceptual understanding, science standards (for example, American Association for the Advancement of Science 1993, NRC 1996, 2000) also emphasize the importance of inquiry skills. Inquiry and conceptual understanding in science often bootstrap one another (Schauble 1990, 1996, Schauble, Glaser, Raghaven and Reiner 1992) and so inquiry is best assessed in a ‘meaningful’ conceptual domain. For this aspect of the evaluation, we selected a curriculum-neutral performance task that involved the creation or revision of knowledge based on evidence that students generate and interpret. The balance-scale task (an adaptation of a training task used by McDermott (1996)) is an ideal task for observing scientific reasoning at the middle-school level. On the surface, the balance scale is a fairly simple apparatus, yet college students typically cannot make accurate predictions about asymmetric configurations without
prior knowledge or instruction (for example, Aoki 1991, Siegler 1981). With extended investigation, however, even younger children can induce the rules for balance (for example, Kliman 1987). The addition of a distractor variable that plays on a common misconception makes it possible to observe how students deal with prior conceptions during the course of experimentation. Moreover, it is a performance task on which students can make progress over the course of one hour.

Objectives of the present study

In this article, results of the in-depth interviews with sixth-grade MARS and non-MARS students are reported. The primary objective of this interview was to evaluate and compare performance on the balance task – a curriculum-neutral performance task (i.e. with respect to content knowledge, neither class covered the concepts of balance or torque). The interview format enables us to examine students’ thinking as it occurs. We were interested in exploring whether, and in what ways, the MARS curriculum might foster students’ scientific reasoning, specifically, their approaches to coordinating experimentation, theory, and evidence. That is, we are interested in both connotations of ‘scientific reasoning’: (a) reasoning that is focused on the understanding of particular scientific concepts or phenomena, and (b) domain-general inquiry and experimentation skills that transcend the particular concepts to which they are being applied (for a review, see Zimmerman 2000).

Individual differences in approaches to scientific inquiry have been documented for children and adults (for example, Dunbar and Klahr 1989, Schauble, Klopfer and Raghavan 1991, Stanovich 1999), and these studies will be used as a guide for analyzing and comparing students’ performance vis-à-vis the coordination of theory, experimentation, and evidence. Previous research has shown that there are two broad categories of strategies used by both children and adults (and indeed working scientists) when performing inquiry and experimentation tasks (for example, Dunbar and Klahr 1989). Discovery tasks that require experimentation in a relatively novel domain can be approached as either ‘experimenters’ or ‘theorists’. These two approaches differ primarily in the manner in which evidence is coordinated with theoretical claims.

Characterizing students as consistent with scientific reasoning profiles is an additional way to look for performance differences between the two groups of students because it is directly relevant to the way students approach a novel task with respect to both inquiry skills and conceptual understanding. Both groups of students were enrolled in innovative science curricula and, on the large-scale assessment, the larger populations from which both groups were sampled performed better than national and/or international averages on items for which such comparisons were available (Raghavan et al. 2002). Therefore, both groups would be expected to perform well, but characterizations derived from qualitative interview data could allow us to explore the nature of any differences in student performance and how these differences may be related to the different instantiations of curricular innovation.

In summary, the present study was designed to explore the impact of the MARS curriculum on students’ reasoning. The performance of students in the MARS program will be compared with those in a standards-based science program attending the same school to identify differences in students’ approaches to scientific inquiry.
Method

Students

Interviews with 14 sixth-grade students (seven MARS and seven non-MARS) from the same school were conducted by a researcher uninvolved with the project. Students were selected by their teachers to provide diversity in terms of race, gender, and scholastic ability.

The school attended by both groups of students is one of several schools participating in the Allegheny Schools Science Education and Technology (ASSET) project. The ASSET project is a local systemic change initiative with the goal of improving science learning and teaching through teacher enhancement and professional development. The ASSET program promotes and trains teachers in the use of nationally recognized curriculum modules including the Science and Technology for Children (STC) series from National Science Resource Center, and the Full Option Science System.

This location was selected as it provided a unique opportunity to sample both MARS and non-MARS students from the same school. Moreover, our goal was to find a comparison group of students – not to exaggerate differences between them and the MARS students. Rather, this comparison group allows an examination of the ways in which MARS students either excelled, or were as good as, students in an inquiry-based science program.

One teacher in the school used the MARS curriculum. The other two teachers taught four hands-on STC modules,2 ‘Magnets and Motors’ (students learn about magnetism, electricity, and electromagnetism and then experiment with different types of motors),3 ‘Measuring Time’ (students investigate how to measure time via moon cycles, sun dials, and water clocks), ‘Experiments with Plants’ (students learn inquiry skills by conducting experiments with plants), and ‘Technology of Paper’ (students learn how to make paper and test it for different properties such as water-resistance and tear-resistance).

Task and apparatus

Students participated in a curriculum-neutral performance task (the balance task), an adaptation of a training task used by McDermott (1996). The balance had three variables that students could manipulate and test (see figure 1). The first variable was the number of weights (washers) that could be placed on hooks on each side of the fulcrum. The second variable was the horizontal position (i.e. the location of the washers closer to or farther away from the fulcrum). The placement of washers along the hooks (numbered 1–6) influences balance, such that the number of weights can be multiplied by the number of distance units to calculate torque. The third variable, vertical position or row (i.e. the location of the washers up or down on the three rows of hooks), is a distractor and does not have an impact on the balance apparatus. Both children and adults, however, have naïve conceptions that lower weights exert more pull or force (Aoki 1991). That is, some individuals believe that a weight placed on the bottom row of the beam ‘weighs more’ than an equal weight placed on the middle or top row of the beam (with horizontal position kept constant). The apparatus had a bubble level-indicator attached to the beam above the fulcrum.
Structured interview and scoring

Prior to the interview, a warm-up task was used to familiarize students with the apparatus. Students were instructed to ‘think aloud’ while interacting with the apparatus as the interviews were both audiotaped and videotaped. The interview itself was composed of six parts and took approximately one hour, resulting in 19–40 pages of transcribed protocol per student.

Variable definition. Students were asked ‘Do you understand what a variable is?’. We noted whether or not the student could respond correctly, with or without prompting.

Variable identification. Students were asked: ‘Can you think of all the different variables that might affect whether the beam will balance?’. We noted the number of focal variables identified (horizontal position, vertical position, and number of washers) without prompting, and the mention of any other reasonable variables. Each time students mentioned a variable, they were asked if they could think of any other variables that might affect the balance. If a student could not mention all three focal variables through this line of questioning, prompts were used to draw attention to the neglected variable(s). These prompts ensured that the students who could not identify all three relevant variables on their own were aware of them before moving on to the variable testing portion of the interview.

Variable testing. For each of the three focal variables, students were asked ‘Is this variable going to affect balance? Can you test it?’. We took note of whether students expressed any expectation or prediction prior to testing (or during variable identification). Performance scores indicated whether students conducted a controlled test of each focal variable (i.e. whether they manipulated the variable of interest while holding the other two constant), and whether or not they correctly assessed the outcome of each experiment. Theory change and/or theory revision was also noted after each test.

Figure 1. Depiction of the apparatus used in the balance task. Washers may be hung from equidistant hooks on either the horizontal or vertical dimension. A bubble level-indicator is attached to the beam above the fulcrum.
**Focused exploration questions.** Focused exploration questions (or ‘focus questions’ for short) allowed the students to investigate the balance, but in a somewhat constrained manner. For the first focus question, one washer was placed on the third hook of the bottom row on each side of the fulcrum. Students were asked to predict what would happen if the washer on the right side were moved to the fourth hook. They were then asked to generate as many ways as they could to make the beam balance by changing the weights and locations on the left side only. There are five possible distinct (i.e. non-redundant) solutions for the first focus question. The number of attempts, the number of successful attempts, and the number of distinct attempts were noted. For the second focus question, one washer was again placed on the third hook of the bottom row on each side of the fulcrum. Students were asked to predict what would happen if a second washer were added to the third hook on the right side. Again, they were asked to generate as many ways as possible to make the beam balance by manipulating only the left side. For this question, there are 11 distinct solutions.

**Rule generation.** Students were asked to come up with a general rule to describe how to make the beam balance. They were free to conduct as many experiments as they wished, without the restriction of keeping one side of the beam fixed (as was the case with the focused exploration phase). They were told that they could use paper to record data. After each experiment, students were prompted with the questions, ‘What is your conclusion from this experiment? Can you come up with a rule?’ The number of additional experiments, if any, was noted. Students’ generated rules were categorized as quantitative or qualitative.

**Application questions.** In the last section of the interview, students were shown four pictures of balances (see appendix 1). Two of the pictures depicted situations with weights stacked on only one peg on each side of the fulcrum. The other two depicted situations with weights distributed across two or three pegs. These questions are a modification of the task used by Stepans (1994). Students were to apply what they had learned from working with the physical apparatus to make predictions about whether the drawings of a different type of apparatus depicted a beam that would tip to the left, tip to the right, or balance. Responses were scored for the number of correct predictions.

**Results and discussion**

The results section will be presented in two parts. Results pertaining to the structured interview questions will be presented first. In the second section, we describe students’ different approaches to the task. Given research suggesting variability in the ways individuals reason about scientific tasks, particularly in coordinating theory and evidence, we outline different profiles described in the literature and then illustrate them with case studies from the verbal protocol data.

For all results reported, coding categories were determined concurrent with the development of the interview questions described earlier. Two individuals coded 28% of the sample (four randomly selected protocols). The initial agreement across coding categories was 95%, and disagreements were resolved through discussion. One individual coded the 10 remaining protocols.
Results pertaining to each section of the interview are described here. Data from the structured interview are presented in table 1. All data in table 1 are reported as a mean percentage for students in the MARS and non-MARS curricula.

Variable definition. After the warm-up activity, the first question that students were asked was to define what a ‘variable’ is. There were no differences between the MARS and non-MARS students with respect to their ability to define a variable correctly, either before or after prompting. Sample responses included simple and succinct definitions, such as ‘It’s something you change’. Some students defined a variable and illustrated with an example relevant to the balance: ‘A variable is something that your – that you can control it or manipulate it . . . Like the uh, washers is a variable and it’s probably going to be manipulated, I’m guessing’. Other students defined a variable using an example from science class to support their definition: ‘A variable is something like, say, like, the plant experiment . . . The different variables could have been like, a soda, a water, no water, or just, like orange juice, something like that. Just like, different things that you’re gonna try testing to see if the experiment – if it makes the experimentation different’.

Variable Identification. Students were asked to look at the apparatus and identify the variables that might affect whether it will tip or balance. There were three critical variables and the number mentioned prior to prompting was noted. Four of the seven MARS students and one non-MARS student mentioned all three variables without prompting. The MARS students, on average, identified 2.43 of the three variables (standard deviation [SD] = 0.79) whereas the non-MARS students identified 1.71 (SD = 0.76), but this difference was not statistically reliable (p = 0.10). A sample protocol of a student who required prompting for clarification: ‘You could change the amount of washers . . . You could change the placement of the washers’. Here ‘placement’ is ambiguous, and the student was asked what she meant: ‘Placement? You could um, maybe put it on the different hooks, ’cause I see that there is three hooks and the rows. . . . Or you could put in on a different

Table 1. Descriptive statistics comparing MARS and non-MARS students on the balance task.

<table>
<thead>
<tr>
<th>Interview component</th>
<th>MARS students</th>
<th>Non-MARS students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable identification</td>
<td>81.0</td>
<td>57.1</td>
</tr>
<tr>
<td>Controlled tests of focal variables</td>
<td>76.3</td>
<td>57.3</td>
</tr>
<tr>
<td>Successful solutions (FQ1)</td>
<td>71.2</td>
<td>39.5**</td>
</tr>
<tr>
<td>Successful solutions (FQ2)</td>
<td>89.5</td>
<td>45.2***</td>
</tr>
<tr>
<td>Application questions</td>
<td>78.6</td>
<td>39.3*</td>
</tr>
</tbody>
</table>

Note: FQ, focused exploration questions. See text for further description of the interview components.

*p < 0.05, **p < 0.01, ***p < 0.001.
number hook. (Recall that the columns of hooks are numbered; see figure 1.) Students mentioned variables other than the three focal variables. The most commonly mentioned variables included the type and/or size of washers used, the type of board the beam was composed of, and type of hooks.

Variable testing. After ensuring that all students were aware of the three focal variables, the interviewer asked the students to test each variable to determine whether it made a difference in whether the beam tips or balances. The MARS students, on average, used controlled tests (i.e. manipulating the variable of interest while holding the other two constant) for 2.29 (SD = 1.11) of three variables, and non-MARS students used controlled tests for 1.71 (SD = 0.95). Although the MARS students were more likely to use controlled tests, the difference was not statistically reliable ($p = 0.32$). As will be discussed later, both curricula included units on variables and variable testing so it is not surprising that performance was comparable for variable definition, identification, and testing.

Focused exploration questions. For the two focus questions (FQ), students were to generate as many ways as possible to make the beam balance by keeping the right side fixed and manipulating weights on the left side only. On average, MARS students generated 5.3 and 8.7 attempts for FQ1 and FQ2, respectively, whereas non-MARS students generated 8.3 and 12.4 attempts for FQ1 and FQ2, respectively. The number of successful solutions was expressed as a percentage of the number of attempts (see table 1). The MARS students outperformed the non-MARS students on this measure for both focus questions ($p < 0.01$). Using this index, the performance of the MARS students appeared to be more systematic, whereas the performance of non-MARS students was consistent with a ‘try-and-see’ approach. That is, if exploration behavior can be characterized along a continuum between ‘random’ and ‘systematic’, an individual taking a ‘try-and-see’ approach would be less likely to generate successful solutions as a function of the number of tries. If, however, one was generating successful solutions for the majority of tries, that would indicate some systematicity, or an understanding of how the apparatus works. Overall, the non-MARS students searched more of the problem space (i.e. they made more attempts), but the MARS students’ exploration of the problem space was more constrained in that they were more likely to narrow in on potentially successful and distinct solutions.

Rule generation. After actively exploring and experimenting with the apparatus during the focus questions, students were asked to come up with a ‘general rule’ that could be used to describe the operation of the balance. They were given the option of conducting as many additional experiments as necessary before answering. There was no difference in the mean number of additional experiments the students carried out.

All seven MARS students were able to express a quantitative rule, although none explicitly stated that one should multiply the mass by the distance on each side of the fulcrum and compare the products. Quantitative rules made reference to information about the number of washers and the number of the hook (i.e. distance from the fulcrum). For example, one student’s rule was ‘You have to have an equal amount of weight on each side, like if you have 2 weights on the 3, that makes 6, so you have to have 6 on the other side’. A second MARS student succinctly stated the
following rule: ‘Just add what – add the left and make it equal to the right – it’s like, multiplying out’. Supporting evidence for the quantitative nature of rules was available during the application questions. Students, either verbally or on paper, would multiply the number of washers (mass) by the position number (distance) on each side and compare the products.

Only one of the seven non-MARS student generated a quantitative rule. The rules generated by four of the non-MARS students to describe the relationship between weight and distance were qualitative in nature. For example: ‘Always try to have the same amount of washers on one side, that can make it even . . . And sometimes you can use one more washer. I keep my extra washer in the inside just to keep it balanced, so try to keep the extra washer near the inside’. These rules made reference to both the weight and distance dimensions, but in a less than precise way: ‘If you have a heavier weight on one side and it’s closer to the board [fulcrum], you should move the lighter weight more and more to the right’. Students expressing rules classified as qualitative had an emerging understanding that both variables are important but they were uncertain of, or could not articulate, the precise nature of this relation. One student was unable to articulate his rule, and one student expressed her rule by describing only symmetrical configurations: ‘Put the same amount of washers in the same place and it would be even’.

Application questions. The interview ended with four paper-and-pencil questions (see appendix 1). Students were shown drawings of an equal-arm balance with upright, equidistant pegs and blocks to represent weights. Students were asked to predict whether the beam would tip to the left, tip to the right, or balance. The MARS students scored 78.6% on the application questions, compared with 39.3% for the non-MARS students ($t(12) = 2.14, p = 0.05$).

The success of the MARS students on the application questions was due, at least in part, to their ability to apply the quantitative rule they had generated to describe the operation of the balance. This was evident in the fact that the MARS students were more likely to solve problems mathematically by either stating out loud or writing below the figures, the mass multiplied by the distance (and sums of products when appropriate) for each side of the fulcrum. That is, it is not the case that MARS students were simply better at multiplying – there was clear evidence that these students discovered and understood that a multiplicative principle underlies the operation of the apparatus. Students who only had a qualitative sense of the relationship between mass and distance were less likely to make correct predictions for asymmetric configurations.

Individual difference profiles

Individual differences or styles in approaching scientific reasoning tasks have been noted by other researchers (for example, Dunbar and Klahr 1989, Klahr and Dunbar 1988, Schauble et al. 1991, Stanovich 1999). As part of our exploration in analyzing the performance of MARS students, we were interested to see whether there were obvious differences in the way students approached this hands-on scientific reasoning task, and if so, whether they might be related to or supported by the different instruction received by the two groups of students.

The individual difference profiles that emerged matched descriptions reported in the literature (for example, Klahr and Dunbar 1988). The students we interviewed
could be described as conforming to three profiles (see table 2): theory-modifying, theory-preserving, and theory-generating. These profiles will be described in and illustrated with selections from the verbal protocol data. Table 3 presents the frequency of profile by curriculum. These profiles are differentiated by certain performance indicators, including the tendency to spontaneously state expectations (or not) about the effect of variables, the tendency to consistently use controlled tests, and the manner in which evidence is evaluated and conclusions are drawn.

Profile: theory modifying

Theory-modifying students spontaneously and explicitly stated their expectations about the role of the different variables in determining whether the beam would balance. Most students had correct intuitions about the role of weight (mass), but held both correct and incorrect conceptions about the horizontal and vertical placement of the washers. The approach to experimentation was mixed – students used both controlled tests and ‘try-and-see’ approaches. As such, the manner in which they evaluated evidence depended on whether the test was confounded or not, and whether they realized the confound or not. When the results of a controlled test were correctly assessed, the student’s theory about the role of the variable was updated to match the pattern of evidence. For example, a student who believed that the horizontal dimension did not make a difference but found evidence that it did, would refine his/her theory about that variable.

<table>
<thead>
<tr>
<th>Profile</th>
<th>Approach to experimentation</th>
<th>Evidence evaluation</th>
<th>Theory/evidence relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory-modifying</td>
<td>Mixed (controlled and try-and-see)</td>
<td>Mixed</td>
<td>Theory revised based on evidence</td>
</tr>
<tr>
<td>Theory-preserving</td>
<td>Mixed (controlled and try-and-see)</td>
<td>Mixed (correct when theory matches)</td>
<td>Evidence distorted to match theory</td>
</tr>
<tr>
<td>Theory-generating</td>
<td>Controlled tests</td>
<td>Correct</td>
<td>Rule derived from evidence</td>
</tr>
</tbody>
</table>

Note: ‘Try-and-see’ refers to relatively free-form exploration of the apparatus.

Table 3. Number of students fitting each profile for each curriculum.

<table>
<thead>
<tr>
<th>Profile</th>
<th>MARS students $\text{n} = 7$</th>
<th>Non-MARS students $\text{n} = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory-modifying</td>
<td>2*</td>
<td>4</td>
</tr>
<tr>
<td>Theory-preserving</td>
<td>1*</td>
<td>2</td>
</tr>
<tr>
<td>Theory-generating</td>
<td>4*</td>
<td>1*</td>
</tr>
</tbody>
</table>

* Students generating the correct quantitative rule.
Case study: Karen
Karen was a non-MARS student whose approach typifies the theory-modifying profile. Samples from her protocol illustrate the key components of this approach to scientific reasoning tasks.

Stated theoretical expectations. During the variable identification section of the interview, Karen pointed out that the horizontal position of the washers was a variable that could be changed: ‘I was just thinking that if you had it inside, it wouldn’t tilt as much’. Karen initially did not notice the placement of weights on the vertical rows as a variable that could be manipulated. Prompting by the interviewer was necessary to draw her attention to the rows: ‘Do you think that might be a variable that you would consider?’ Karen replied ‘I guess, because if you put one here [pointing to the bottom row], I think it would go – like, go down quicker than if you would put one up here [pointing to the top row]’. She acknowledged that the variable could be manipulated, and concurrently expressed a hypothesis about its role in balance.

Experimental approach and evidence evaluation. Karen had a mixed approach to experimentation and her evaluation of evidence. For example, during the variable testing portion of the interview Karen was asked whether she could test whether the number of washers made a difference in whether the beam tips or balances. The balance had one washer on each side in the third position, bottom row. She initially performed a confounded test by adding an extra washer to the left side on the fourth position, in the same row. After a prompt asking if that test showed whether the number of washers made a difference, she said ‘I don’t know . . . you could try to even out the, like, washers in a way’ and then produced a symmetrical (i.e. balanced) configuration. She proceeded in an exploratory manner. She would add one washer to the left side, then add a washer to the corresponding symmetrical location on the right side a number of times to produce balanced situations. After several prompts she did not produce a controlled test of the weight variable. When asked whether the number of washers made a difference, she said: ‘Yeah, it’s – because like, if you have eight washers on this side and only three washers on this side, since it has more washers and washers are like heavier, that means that this side is gonna be heavier’.

The previous paragraph illustrates that Karen was not using controlled tests, but was using prior knowledge rather than experimental results to draw a conclusion about the weight variable. To contrast, the following interaction shows that Karen does know how to describe and perform a controlled test:

Researcher: Is there a test that you can do to show whether or not the washer is on the first row or the second row or the third row makes a difference in whether or not it’s tipped or balanced?
Karen: Probably. Because if you keep both of them in the same row, like, say this is the column you wanted to work in. So column number 3. You have to make sure it’s in column number 3 the whole time, so say you put one washer on the bottom and one washer on the top, you can see which one would have more weight.
Researcher: Okay. So, how is it looking so far? Where is the bubble?
Karen: It looks like the bubble is like, closer to this side.
Researcher: A little closer to that side?
Karen: Or it looks even, in a way.
Researcher: So then from this test, can you tell whether or not the height –
Karen: Mmm, not really. I don’t think the height really matters . . . Well, usually when you would think of it, you would think that it would be on the bottom or else it would pull it – the gravity would – it would pull it down faster, but now I’m not sure, and it still looks kind of even . . . So I guess that’s usually what happens and like, the height – I really don’t think matters since we had it in the same [column].

Karen describes a controlled test of the vertical placement, but seems to expect that the beam will tip in stating that one side would have ‘more weight’. Given the outcome, however, she tries to resolve the inconsistency between her predicted expectation and the evidence she is evaluating. As she talks it through, she updates her theory by accepting that vertical placement is not a factor in whether the beam tips or balances.

Final rule and robustness. Karen’s general rule was used earlier to illustrate the nature of students’ qualitative rules. Given that her rule was qualitative, she made incorrect predictions for three of the four application questions. The qualitative rule ‘works’ for the first question, but for the more complex questions it is insufficient. She tested her predictions by arranging washers on the balance in the configuration shown in the pictures. Although her predictions were incorrect, she seemed unable to use this pattern of evidence to further modify her qualitative general rule or theory of how the beam balanced. That is, although she knew her rule was not precise enough to make accurate predictions, she was not able to move from her qualitative understanding of the data patterns to discern a more powerful rule.

Profile: theory preserving

The profile of the theory-preserving student in all ways matched that of the theory-modifying student, except for one crucial difference – these students had the tendency to preserve an existing theory about the role of one or more variables despite the evidence. Most frequently, theory preserving centered on the belief that the vertical dimension influenced whether the beam tips or balances. Less frequently, some students believed that the horizontal dimension does not influence the balance. This pattern of behavior was reported previously by Kuhn, Amsel and O’Laughlin (1988), who noted that children would distort the evidence to fit a favored theory.

Case study: Albert

Albert was a student from the MARS curriculum who fit the profile of a theory preserver. Despite preserving an incorrect notion about the role of the vertical dimension, he still generated the correct quantitative rule to describe the operation of the balance.

Stated theoretical expectations. Albert spontaneously explained what he thought the effects of the placement of weights would be. With respect to the horizontal dimension: ‘ . . . as the – the ring [washer] gets closer to the middle, it doesn’t tip as much, like um, if you put a pencil on the ground and you put a ruler over it, or like a seesaw – if you sit in the middle of a seesaw, it doesn’t move. But if you sit on the sides, it moves more’. Albert’s expectation about the effect of the vertical position,
prior to testing this variable was as follows: ‘[T]he bottom probably pulls down more because, uh, the farther you get up, the less board is pulling down. Like if it’s on the bottom, it just pulls down more’.

Experimental approach and evidence evaluation. Due to his belief about the role of the vertical placement in determining whether the beam balances, Albert incorrectly assesses the outcome of a controlled experiment, in which he placed washers in the sixth position on each side, with the left washer on the top row, and the right washer on the bottom row: ‘Well, the one – the one – the lower hooks, um, it bends a little bit towards that way, tips a little towards that way’. The interviewer then directed his attention to location of the bubble in the level-indicator attached to the top of the beam (indicating a balanced situation), to which Albert replies: ‘. . . but it’s just going a little bit out of the black lines. Yeah, it’s a little bit out of the – a little bit – a little while there’.

Albert used controlled tests during the variable testing phase of the interview, and during the focus question tried a number of exploratory configurations that were successful. At one point during a focus question he disregarded the instruction to keep the right side fixed and moved the one washer from the fourth position to the sixth position and put two washers on the third hook on the left side: ‘I don’t know how I thought of this, but I just like, since I figured three and one, you add it to four, so I’m guessing if you put it on the six, if you put two on the three it might balance. Yeah it’s balanced’.

Although he began to notice the numerical relationship between weight and distance from the fulcrum, his first attempts (once he was back on task) were unsuccessful. This is critical because they were his first deviations from using the bottom row only (as seen earlier, he believes the vertical dimension affects the balance). The unsuccessful attempts to create a balanced situation when the right side had one washer on the fourth hook (bottom row) included configuring the left side with one washer on the bottom row in the first position and one washer on the top row in the fourth position (tips to the left). In a second attempt, he moved the washer in the first position to the top row (tips to the left). Albert evaluated these results to be consistent with his theory that the row in which the washers are placed influences the balance.

He then moved one of the weights on the left to the third position, a configuration that resulted in the beam balancing.

**Researcher:** There you’ve got three and one.
**Albert:** Okay, but you should keep it in the same row, because then you can even out the weight more, ‘cause if they’re in different rows . . .

**Researcher:** So they should be in the same row, but what’s the outcome of what you’ve done here? Is it balanced or is it tipping?
**Albert:** Yeah, it’s balanced.

**Researcher:** So does row make a difference then?
**Albert:** Well, it’s – it tips just a little bit this way, but it – so it doesn’t make that much of a difference – not a very noticeable difference, but if you want to be exact, it makes a difference.

**Researcher:** So if you wanted to be exact, you would keep them in the same row?
**Albert:** Yeah.

This passage illustrates that although he can acknowledge that the difference is not a ‘noticeable’ one (i.e. it is balanced), Albert still clings to the idea that the row
makes a difference but that it is just barely perceptible. That is, he incorrectly interpreted the evidence (or misperceived it) as consistent with his theory.

**Final rule and robustness.** For the second focus question, the left side has two weights on the third position (bottom row), and each of Albert’s attempts was both successful and distinct. The following dialog shows that he determined the numerical properties underlying balance:

- **Researcher:** Okay, so you’ve got one on the third, one on the second, and one on the first hook. Why did you do that?
- **Albert:** Because three plus two plus one is six and over here, there’s two rings on the third one. Three plus three is six, and six and six on each side is like algebra.

When Albert was asked to describe a ‘general rule’ to make the beam balance, he continued to focus on the role of the vertical dimension:

- **Albert:** Okay, well, first – I’ll take the washers off first. I’d prove that the – that the lower ones usually pull down more.
- **Researcher:** Mm-hmm.
- **Albert:** You can put one on the same column, only higher and down here on the same column, and it tips this way. I’m not sure why, but I just remember something about science class that if – as long as it’s lower, it’ll weigh more.
- **Researcher:** Yeah? Lower things weigh more? Okay.
- **Albert:** They’re closer to the center of the earth, where the pull from gravity is.7
- **Researcher:** Okay, and so what is the outcome of this experiment that you’ve done?
- **Albert:** It’s pretty even, but the – it tips a little left and uh, about the adding, so you keep one on six, you put two on three and that balances because three plus three is six, and six on the other side.
- **Researcher:** Okay, so your general rule would be . . .
- **Albert:** That um, if you a – if the positions are exactly the same, say there’s an inch between each uh, hook, and as long as they’re an inch over here, because say that the six hook is over here for this one, but this hook’s right there for this one–
- **Researcher:** Yeah.
- **Albert:** They – it’d make a difference. But as long as they’re in the exact opposite position, um, you can – if there’s six on this side, you can add them up to the six on this side, or four and four over here.

The application questions (see appendix 1) did not include the vertical dimension, and therefore Albert got all four predictions correct. During the course of the interview, he demonstrated that he induced the mathematical rule involving the multiplication of weight and distance from the fulcrum, but he did not change his theory about the importance of the vertical placement.

**Profile: theory generating (experimenters)**

The theory-generating profile most closely matched Klahr and Dunbar’s (1989) description of ‘experimenters’. Theory-generating students did not spontaneously verbalize any expectations about the effect of the variables during the variable testing phase of the interview.8 They usually experimented by using controlled tests to determine the effect of each variable. Moreover, they were systematic in exploring solutions during the two focus questions requiring asymmetric balance configurations. All of the students fitting this profile discovered or induced the correct quantitative rule from the generation and evaluation of evidence (regardless of
curriculum; see table 3). Thus, the name ‘theory generating’ is appropriate because they approached the task without stated preconceived notions but rather they generated the theory (i.e. the rule for balance) based on the pattern of data they obtained during their experimentation.

Case studies: theory generators
The theory-generating profile is not as easy to illustrate with a single case study from our data. Therefore, protocols from a number of students will be used.

Unstated theoretical expectations. The fact that this type of student was reticent with respect to expectations makes it difficult to illustrate with portions of verbal protocol. When verbalizing an expectation, a typical statement was in the form of ‘I will see what happens’ rather than ‘I expect X to happen’. Marjorie, a MARS student, describes how to test whether the vertical placement makes a difference: ‘Well, you need the same amount on each side, since we’re doing the same, and you could move this one to the top and see if that made a difference’ (emphasis added). That is, theory generators take the stance of withholding judgment until after the evidence is produced: ‘You could try moving one of the [washers] and see what happens then . . . like, moving like, it either up or to the sides’.

Experimental approach and evidence evaluation. Marie is a MARS student who exemplifies the theory-generating profile with respect to the use of controlled tests over ‘try-and-see’ exploration. Given the tendency to prefer controlled tests, these students assessed evidence correctly and used it in the generation of a theory about the balance beam.

Marie: So maybe if you manipulate it by putting – controlling this row by putting that one just on the three and maybe moving this one to four.
Researcher: Okay, and so which variable are you testing there?
Marie: Um, the movement of the washers.
Researcher: Okay.
Marie: I would be controlling one side and manipulating the other.
Researcher: Okay, and so by moving that to the fourth hook – does that affect the balance?
Marie: It affects the balance.

Final rule and robustness. As stated previously, all theory-generating students produced the correct quantitative rule, regardless of curriculum (four MARS students, one non-MARS student). Except for one student, the theory-generators correctly applied their quantitative rule to the application questions. The one exception, Ginny, despite a correctly stated rule, failed to apply it in the case of three out of four questions. After the last question she tried the configuration on the balance to test her prediction, and it was wrong. This prompted her to re-evaluate all four questions by calculating the products on each side and comparing them, testing the configuration on the balance, which then solidified the correctness of her rule.

Profiles and curriculum
The number of students fitting each profile for each curriculum is presented in table 3. Due to low expected frequencies, it is not possible to determine whether profile and curriculum are statistically independent or not. An interesting pattern does
emerge, however, in that the MARS students were successful in generating the correct quantitative rule, regardless of profile. Only one student from the non-MARS group was classified as an experimenter, and she was the only student from that group to generate the correct quantitative rule. The rest of the non-MARS group were classified as theorists (theory-persevering and theory-modifying) and these students did not induce the correct rule for balance. The relationship between scientific-reasoning profile and curriculum will be discussed in more detail later.

General discussion

In the current study, we explored whether or not students in a science program emphasizing model-assisted reasoning performed as well as those in a nationally recognized program on a curriculum-neutral performance task. The present study represents an initial attempt to document the performance of small groups of MARS and non-MARS students during in-depth interviews in which students experimented with a novel inquiry-based reasoning task. The goal was to determine whether, and in what ways, students’ approaches to the balance task differed and whether they could be related to science instruction. Given that some differences were evident, we begin by summarizing the results and then discuss key features of instruction in each curriculum that might account for the patterns of performance that were observed.

Both groups did well on the curriculum-neutral task, displaying basic inquiry skills such as defining, identifying and testing variables. There were, however, ways in which the MARS students excelled. (a) MARS students generated more unique and successful configurations of asymmetric balance during focused exploration. (b) All MARS students, regardless of approach to inquiry (i.e. classification as experimenters or theorists), were able to discover a quantitative rule for determining balance. All but one of the non-MARS students generated a qualitative rule to account for the interaction between weight and horizontal position. The discovery of the correct rule for determining balance is especially significant given that college students typically do not know the torque rule unless formally taught it (for example, Siegler 1981, Chletsos, DeLisi, Turner and McGillicuddy-DeLisi 1989, Aoki 1991). (c) MARS students were more likely to take an experimental approach, using the patterns in the data to induce a mathematical rule rather than approaching the task with explicit predictions. (d) MARS students were more successful in applying what they learned with the physical apparatus to solve related paper-and-pencil problem situations.

One goal of the present research was to determine whether, and in what ways, the MARS program fosters the use of scientific process skills given that a key emphasis of the curriculum is on the use of conceptual models. As mentioned previously, research to date on the MARS program (for example, Raghavan et al. 1998a, 1998b) focused on documenting students’ conceptual change (e.g. concepts of mass, volume, density, buoyancy) for the purposes of curriculum development. Previous work demonstrated that model-centered instruction can help students understand science concepts, but it is also important to document the ways in which students learn and apply scientific reasoning skills. In the next sections, we describe and compare typical lessons students in each group were exposed to in order to shed light on how curriculum differences may have fostered the differences in performance and approaches taken by the MARS and non-MARS students. We
illustrate the way experimentation is taught to the non-MARS students using STC modules, and the type of lessons that incorporate experimentation in a MARS module, followed by a critical comparison.

*An example of experimentation in the ASSET curriculum*

Experimentation skills are emphasized throughout the sixth-grade curriculum, but typically are incorporated within units in which conceptual knowledge is also important. For example, students in the ASSET program are introduced to experimentation in the STC module, ‘Experiments with Plants’. The main objective of this 8-week module is ‘to teach students how to design and conduct controlled investigative experiments’ (National Science Resource Center 1992: 1). Students first learn about the key variables that influence the health and reproductive capabilities of Wisconsin Fast Plants®. Once they identify the variables that affect plant growth, students work in teams to formulate a question to answer by manipulating one variable and holding others constant. Teams set up experimental and control plants and collect data on a daily basis. While plants are growing, students learn more about, for example, plant anatomy, pollination, and germination. After several weeks of observations and measurements, teams draw conclusions from their cumulative data and decide whether the investigation answered the experimental question.

*An example of experimentation in the MARS curriculum*

Key features of the MARS curriculum were described in detail in the Introduction. The 2-day lesson, ‘Compare Lumps of Clay’, in the ‘Mass & Density Unit’ is an example of a type of hands-on experimentation in which students frequently engage. Students working in pairs are given a lump of modeling clay and asked to predict what will happen to the volume, mass, and density when the lump is cut in half. After recording their predictions, students measure the mass (in grams) and the volume (in milliliters) and calculate the density of the whole lump, and design a data table to display their measurements and calculations (mass, volume, density). They repeat the procedures for one-half, one-quarter, and one-eighth of the lump. Sixth graders typically expect all three quantities to be reduced by one-half, and they are generally surprised to discover that the density values change very little. By examining the patterns in their data, students can conclude that, unlike volume and mass, which are properties of objects, density is a property of the material the objects are made of; in this case, clay.

Due to measurement error and rounding, calculations for the density for each lump will be off by a decimal place or two, but not by one-half, which is the predicted outcome. This pattern of decreasing by one-half, however, is found in the measurements for the volume and mass of each lump of clay. Because density values vary somewhat from lump to lump but come nowhere close to decreasing by half each time, this particular lesson also helps students learn to differentiate between errors in data and patterns in data. Throughout the sixth-grade module, students are engaged in similar activities, recording predictions, generating data, devising some type of data model to display their results, examining the data model for patterns, comparing predicted and actual outcomes, and drawing conclusions about underlying concepts.
Curriculum comparison

Both curricula emphasize process skills such as asking questions, collecting data, and drawing conclusions. The comparison between two activities already described highlights important differences in the way students learn about experimentation. In the STC module, students collect data on plants for an extended period, and after several weeks are to draw a conclusion based on the cumulative data. In some instances, it is possible for a team of students to collect data over the course of several weeks, only to find that the data are ambiguous or do not lead to a conclusion. As sometimes happens in science, results may be unclear, but in the case of a plants experiment there is no time to conduct a ‘do over’ experiment. Over the course of several weeks, in contrast, MARS students collect and analyze multiple sets of data, and experiments are designed such that the data generated will reveal patterns that enable students to draw conclusions. Therefore, MARS students encounter data more frequently and that data are usually of better quality with respect to their ability to support conclusions. In the time it takes to run a single plant experiment, MARS students encounter multiple opportunities to design and analyze data models, and thus gain more experience at making sense of numbers and interpreting experimental results. This may explain why MARS students were better able to formulate a quantitative rule for the balance than were non-MARS students.

During the balance task, MARS students were less likely than non-MARS students to make spontaneous predictions prior to collecting data. This might be due to the fact that MARS students perform a number of experiments that yield results that contradict their expectations. For example, MARS activities are designed to address such common misconceptions as density decreases with volume and mass, thick (viscous) liquids are denser than thin liquids, and heavy objects displace more water than light objects. After a number of experiences in which their predictions are not supported by data, students may be inclined to withhold judgment about a particular variable until after the data are collected and the evidence is analyzed.

Almost all non-MARS students were classified as theorists. One key feature that distinguishes theorists from experimenters is the tendency of theorists to begin with a hypothesis (e.g. by speculating about the role of different variables prior to empirical testing). We do not wish to make evaluative judgments about whether one approach is superior to the other, given that within the scientific community individual scientists claim to specialize as either experimenters or theorists (Simon 1986, Bauer 1992, Klahr and Carver 1995). Taking a theoretical approach with a novel scientific reasoning task may be more challenging, given that belief revision has been shown to be more difficult than forming new theories when prior beliefs do not exist or are not held with conviction (for example, Holland, Holyoak, Nisbett and Thagard 1986, Koslowski 1996, Ruffman, Perner, Olson and Doherty 1993). As seen with the MARS students, however, it is possible to take a theoretical approach and still discover a quantitative rule for balance. The MARS students, however, received instruction and practice in coordinating evidence with theory in the successive refinement of data models. That is, model revision is an integral part of the learning sequence, and instruction necessitates building upon previous knowledge. Without multiple opportunities to draw conclusions and refine models, in contrast, non-MARS students who took a theoretical approach were less successful on this type of scientific reasoning task.
Summary

The interview data presented represent an initial attempt to document and describe differences in students’ experimentation strategies, and how they relate to MARS instruction. Differences between students from two different curricula were evident on a novel performance task involving the coordination of theory and evidence. Most striking was the number of students who generated the correct quantitative rule. All MARS students discovered this numerical relationship through experimentation, regardless of their scientific reasoning profile (i.e. experimenters or theorists). To make sense of these data, we reflected on the critical differences in the types of instruction in the two science curricula. Both MARS and non-MARS students were involved in extended investigations of scientific phenomena. Three differences were apparent: (a) MARS students had multiple opportunities to collect data and draw conclusions about the same topic, whereas the non-MARS students collected data over longer periods of time with limited opportunities to draw conclusions; (b) MARS lessons were designed such that the results of experimentation frequently contradict predicted expectations; and (c) the multiple opportunities that MARS students had to draw conclusions required successive refinement of models. This emphasis on model revision, rather than just model generation, may be the crucial component of instruction that allows students to develop skills to flexibly coordinate theory and evidence.

Acknowledgements

This research was supported by a grant from the Office of Educational Research & Improvement, US Department of Education. The authors thank Laura Moin for assistance with transcription and coding of videotapes.

Notes

1. Although there are many definitions of what counts as ‘theory’, we will use the term in a manner consistent with that used by Kuhn and Pearsall (2000: 116–117). They outline four possible uses of the term ‘theory’ or ‘theoretical claim’, which range from least stringent such as category and event claims (e.g. ‘this plant died’) to most stringent such as causal or explanatory claims that include an explanation of why the claim is correct (e.g. ‘this plant died because of inadequate sunlight’). The commonality among theoretical claim types is that ‘although they differ in complexity, each . . . is potentially falsifiable by empirical evidence’ (Kuhn and Pearsall 2000: 117). At the middle-school level, much of the theoretical claims that students deal with in pedagogical or assessment tasks are less stringent than explanatory system claims, although they may learn these types of theories in a more didactic way (e.g. the theory of evolution, the kinetic theory of gases).
2. For more information, see www.si.edu/nsrc/pubs/stc/matrix.htm
3. MARS students also received instruction in the ‘Magnets and Motors’ module prior to the MARS units.
4. The precise question was to generate ‘as many ways as possible . . .’. A student who knows that the vertical dimension does not matter may disregard it and focus on distinct (non-redundant) solutions, disregarding row. We counted the distinct solutions and expressed this as a percentage of the number of attempts. The pattern was similar to that obtained with successful solutions, with percentages being smaller than successful solutions for both groups, but with the MARS students outperforming the non-MARS students on both focus questions (p < 0.05).
5. Torque is calculated by multiplying mass by distance, but we would expect students to substitute the number of washers for mass, and the position number for distance from the fulcrum.
6. Students’ names have been changed.
7. Albert seems to be misapplying the concept of gravitational pull. Although theoretically correct, the difference in height between the rows is negligible to practically make a difference.
8. That is, unstated theoretical expectations were part of the behavioral profile. This is not to say that they did not have them. They simply did not report predictions or hypotheses except when prompted by interview questions.

References


Appendix 1  Application questions

Figure 1

Figure 2

Figure 3

Figure 4