Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students

Barbara Y. White
Graduate School of Education
University of California at Berkeley

John R. Frederiksen
Division of Cognitive and Instructional Sciences
Educational Testing Service
Oakland, California

Our objective has been to develop an instructional theory and corresponding curricular materials that make scientific inquiry accessible to a wide range of students, including younger and lower achieving students. We hypothesized that this could be achieved by recognizing the importance of metacognition and creating an instructional approach that develops students’ metacognitive knowledge and skills through a process of scaffolded inquiry, reflection, and generalization. Toward this end, we collaborated with teachers to create a computer enhanced, middle school science curriculum that engages students in learning about and reflecting on the processes of scientific inquiry as they construct increasingly complex models of force and motion phenomena. The resulting ThinkerTools Inquiry Curriculum centers around a metacognitive model of research, called the Inquiry Cycle, and a metacognitive process, called Reflective Assessment, in which students reflect on their own and each other’s inquiry.

In this article, we report on instructional trials of the curriculum by teachers in urban classrooms, including a controlled comparison to determine the impact of including or not including the Reflective Assessment Process. Overall, the curriculum
proved successful and students’ performance improved significantly on both physics and inquiry assessments. The controlled comparison revealed that students’ learning was greatly facilitated by Reflective Assessment. Furthermore, adding this metacognitive process to the curriculum was particularly beneficial for low-achieving students: Performance on their research projects and inquiry tests was significantly closer to that of high-achieving students than was the case in the control classes. Thus, this approach has the valuable effect of reducing the educational disadvantage of low-achieving students while also being beneficial for high-achieving students. We argue that these findings have strong implications for what such metacognitively focused, inquiry-oriented curricula can accomplish, particularly in urban school settings in which there are many disadvantaged students.

Many cognitive science and educational researchers theorize that much of learning is a conscious process that incorporates metacognitive knowledge and skills as key components (A. Brown, Bransford, Ferrara, & Campione, 1983; Collins, Brown, & Newman, 1989; Feuerstein, Jensen, Hoffman, & Rand, 1985; Flavell, 1979; Nickerson, Perkins, & Smith, 1985; Resnick, 1987; Wellman, 1984). These metacognitive skills and knowledge can be acquired, and so, the argument goes, students can “learn how to learn.” Creating curricula that help students to develop an awareness of their inquiry process and an ability to reflect on it could enable students to improve their learning expertise while also acquiring subject matter expertise. A focus on developing such learning and inquiry expertise is missing in most school curricula.

In the research we report here, we set about to apply such a metacognitive view of learning to science education and to evaluate its utility. To achieve this goal, we first created a simple but general model of the scientific inquiry and learning process. We did this by transforming the steps in the instructional cycle developed in our prior research on science education (B. White, 1993b; B. White & Horwitz, 1988) into an Inquiry Cycle that is explicitly presented to students (as shown in Figure 1). In following this cycle, students pursue a sequence of research goals in which they first formulate a question and then generate a set of competing predictions and hypotheses related to that question. In order to determine which of their competing hypotheses is accurate, they then plan and carry out experiments (using both computer models and real-world materials). Next, they analyze their data and summarize their findings in the form of scientific laws and models. Finally, they apply their laws and models to various situations. As they do this, they reflect on both the limitations of what they have learned (which suggests new questions) and on the deficiencies in the inquiry process itself (which suggests how it could be improved). Thus, after this reflective process, students are back at the beginning of the cycle with a new or refined question and a revised approach to inquiry.

The view of inquiry embedded in this cycle is quite authentic in that modern views of scientific methodology map readily onto this constructivist, metacognitive
view of learning and inquiry (Chalmers, 1990; Giere, 1991). The ThinkerTools Inquiry Curriculum was designed to facilitate the learning of this inquiry process as well as learning how to monitor and reflect on it. It engages middle school students in authentic scientific inquiry in which their primary goal is to create and apply causal models of force and motion. This contrasts with many science curricula that focus on memorizing definitions, facts, and formulas and on learning to solve stereotypical problems via the application of formulas using acausal, algebraic reasoning. Thus, in our science curriculum, the purpose is to enable students to learn about the process of scientific inquiry and modeling while at the same time learning about the physics of force and motion. The subject matter is thus physics, but because the focus is on inquiry, it could just as well be biology or any of the other scientific disciplines.

In creating this curriculum, we were particularly interested in counteracting the view that science is an abstract and difficult discipline that is accessible to only an elite subset of the population—namely, to high-achieving students over the age of 13 who, it is argued, are the only ones capable of the abstract, complex reasoning processes needed to learn and do science. This view is widely held, particularly in the case of physics, which is often regarded as the most abstract and difficult of all the sciences. The serious study of science is regarded as beyond the reach of most elementary and middle school students, particularly in urban classrooms in which many students are low achievers.

Based on research revealing the importance of metacognition, our hypothesis is that the reason students have difficulty with science, particularly physics, is not that they are too young or lack intelligence, but rather that they simply do not know how to construct conceptual models of scientific phenomena and how to monitor and reflect on their progress (cf. Campione, 1987; Carey, 1985; Feuerstein et al., 1985; Nickerson, et al., 1985; R. White & Gunstone, 1989). Thus, if you teach
students about the processes of scientific inquiry and modeling, and if you also teach them how to monitor and reflect on their inquiry processes, then they can engage in inquiry and learn physics as well as older or higher achieving students can.

Thus, we argue the need for curricula designed to scaffold the development of students’ inquiry, modeling, and metacognitive skills and knowledge. The scaffolded development of such knowledge should be helpful to all students, particularly the younger and the lower achieving students because this knowledge is a crucial part of the conscious expertise about learning that they lack. Furthermore, enabling students to acquire such expertise should reduce the differences in learning between younger and older students and between lower and higher achieving students.

To test these hypotheses and to assess the effectiveness of the ThinkerTools Inquiry Curriculum in achieving the goal of making science accessible to a wide range of students, three teachers implemented the curriculum in a total of 12 urban classes. We carried out a rigorous evaluation of the students’ inquiry and physics expertise using pretests and posttests. We also evaluated the quality of their research projects, which were undertaken as the curriculum progressed. The major factors we looked at in our various analyses were grade level (7–9) and achievement level (based on scores on a standardized achievement test). This enabled us to determine whether the younger and lower achieving students did as well as older and higher achieving students. We also incorporated a controlled study to investigate the importance of the monitor-and-reflect metacognitive components of the curriculum, which are introduced to students in the form of the Reflective-Assessment Process.

An overview of the article. In this article, we begin by presenting the theoretical bases for our work. We go on to describe the ThinkerTools Inquiry Curriculum: First, we outline our goals, namely the type of expertise that we want students to acquire—both with regard to their inquiry skills and their understanding of the physics. Then, we present an overview of the instructional approach and describe the ThinkerTools software and its role in facilitating scientific inquiry and modeling. Next, we describe how the instructional approach and materials are actually employed in the classroom; here, examples of the inquiry and reflection activities and their sequencing and rationale are given. Following these descriptions of the ThinkerTools Inquiry Curriculum, we present the results of instructional trials in urban classrooms. In particular, we present data regarding its effect on (a) students’ scientific inquiry expertise, (b) their physics knowledge, and (c) their attitudes about learning science. We conclude by discussing implications and challenges raised by our findings.
THE THEORETICAL APPROACH

Inquiry-oriented, constructivist approaches to science education are not new. Since the 1960s, there has been a great deal of research and curriculum development work aimed at teaching the "scientific method." Unfortunately, these earlier efforts met with limited success. Teachers typically found inquiry curricula to be difficult and time consuming to teach, they frequently stated that they lacked the training and supports needed to implement them well, and they believed that such curricula are not effective for their students (Stake & Easley, 1978; Welch, 1981). Furthermore, measures of students' performance indicated that the teachers' perceptions were accurate: Many students were not learning much science and inquiry from these various inquiry-oriented approaches (Harms & Yager, 1980). What has changed that causes us to be more optimistic? In short, our work synthesizes recent advances in cognitive science (particularly the work on metacognition) with advances in educational technology (particularly the creation of computer simulation and modeling tools), which we believe set the stage for developing more effective approaches to the teaching of scientific inquiry. Furthermore, one can create exemplary materials and supporting social structures, such as videotapes and video clubs (Frederiksen, Sipusic, Gamoran, & Wolfe, 1997), which can help teachers to successfully implement such new approaches.

Creating Scientific Models

We start from the theoretical perspective that complex theories in science are developed through a process of successive elaboration and refinement in which scientific models are created and modified to account for new phenomena that are uncovered in exploring a domain. In our prior research, we hypothesized that similar processes are involved in students' learning of scientific concepts and in their development of conceptual models (B. White, 1993a, 1993b; B. White & Frederiksen, 1990; B. White, Frederiksen, & Spoehr, 1993). Our approach was to create a series of experimental situations that require students to make successive refinements to their conceptual model in order to accommodate increasingly complex phenomena. In this research, many of the experimental situations were set in the context of computer models and simulations. These simulations incorporate causal models and representations that depict in iconic, graphic, or symbolic form the important concepts, principles, and processes found in the scientifically accepted theory of the domain (B. White, 1993a). Students who were given this sequence of experimental situations and computer models were able to develop, through successive refinements, increasingly complex and accurate conceptual models for the domain (B. White, 1993b; B. White & Frederiksen, 1990).
In developing these prior curricula, the decisions about which experimental situations were appropriate for students to explore at each stage of learning were made when the curriculum was constructed, using heuristics such as "start with simpler situations and progress to more complex ones." The focus was on optimizing the students' opportunity to construct the requisite conceptual models, not on developing the inquiry skills needed to find appropriate experimental situations to investigate, that is, situations that are likely to be profitable to explore at each stage in learning.

Although this instructional approach based on progressively elaborated models has proven to be effective, it shares with many other science curricula a prescribed sequencing of experimental activities that students must follow. In contrast, in authentic scientific research, investigators must develop for themselves the experimental and observational plans that will enable them to test the adequacy of their current theory and to improve, over time, its generality and predictive power. We therefore set out to develop and test an inquiry-based instructional approach in which students learn how to apply an inquiry process that allows them to generate for themselves new situations and experimental plans that will enable them to test and improve their conceptual models. Students are introduced to a general heuristic for managing their inquiry: mainly, simplification followed by the controlled introduction of complexity. The actual process of carrying out their investigations is represented as a five-step cycle (Question, Predict, Experiment, Model, and Apply). By incorporating this Inquiry Cycle, our approach to science learning is constructivist in two senses: in the students' approach to learning through a process of successive model elaborations, and in their inquiry process used in generating situations and experiments that extend their understanding of nature.

In creating this approach, we also paid close attention to modern accounts of the epistemology of science. We wanted to construct a social environment within the classroom that approximates that of a community of researchers working within a domain of scientific inquiry (Chalmers, 1990; Matthews, 1994). According to this postpositivist view, the community is responsible for developing a consensus about what are the important theoretical concepts to consider, how these concepts are lawfully related within a model, and how such models can be used to represent real-world behavior. The community must also assess the results of experiments and observations they have carried out and judge their relevance and implications for the models they are constructing.

The Instructional Dilemma

To implement this approach to learning and inquiry, we need to create a classroom research community—one that applies an inquiry process to create progressively more adequate models through a principled process of experimentation, model building, and application. This is a complex individual and social activity, one that
is seldom practiced even in university-level science courses, let alone in middle
school classes. The paradox is that, to understand this complex activity, one needs
to do it, but to do it, one needs to understand it. The instructional solution we
developed combines aspects of prior work on metacognition (A. Brown et al., 1983)
and situated cognition (J. Brown, Collins, & Duguid, 1989). It scaffolds carefully
the inquiry process for students so that they can begin practicing it. This is analogous
to the first stage of an apprenticeship in which novice participants enter a commu-
nity of practice (Lave, 1988); however, in our approach we wish to avoid creating
situations in which students begin as peripheral participants (Lave & Wenger,
1991). Rather, we follow the ideas of Reciprocal Teaching (A. Brown & Palincsar,
1989; Palinscar & Brown, 1984) and attempt to create at the outset an activity
sequence that approximates the mature phases of scientific inquiry.

We expect that, initially, students' understanding of inquiry will be only proce-
dural—they will develop what amounts to a knowledge of the set of interrelated
activities. With practice and reflection, however, they will develop a deeper under-
standing of the functions of inquiry (the nature of models and how to develop more
general and more powerful models), of the phases of the Inquiry Cycle (questioning,
predicting, experimenting, modeling, and applying), of the heuristics involved in
linking phases of inquiry, and of the role of social processes in developing and
revising models. Our view is that these aspects of systematic thought cannot be
appreciated by students until they are actually carried out and reflected on. The
instructional process thus seeks to create conditions under which this will occur.

The bridge to understanding this complexity is reflective peer and self-assess-
ment (J. Brown et al., 1989). Participants in the classroom research community
need to be involved explicitly in a reflective process in which they review their
processes of working and the products of their investigation (their research reports).
Furthermore, this reflective process should be a social one so that students may see
how multiple perspectives can be applied in viewing one's own and others' work
as they carry out the process of inquiry and modeling. This social process allows
students to practice and internalize habits of reflection (Vygotsky, 1978). For these
reasons, we have introduced a peer and self-assessment activity into the classroom.
Students are presented with a set of aspects of thought that serve as a set of
"prompts" to encourage them to examine important aspects of their inquiry and
modeling processes. Examples include "Being Systematic" and "Being Inventive." Introducing this Reflective-Assessment Process is, we hypothesize, the key to
building an understanding of inquiry and an ability to carry it out effectively.

In teaching these inquiry and reflective processes, we cannot utilize the cognitive
apprenticeship approach advocated by Collins et al. (1989). In most middle school
science classrooms, there is no preexisting research culture. Also, middle school
science teachers often do not have advanced degrees in science, and those who do
rarely have been members of research communities. We do not, therefore, have
available expert practitioners of inquiry-based research who can serve as models
for students, as in a traditional apprenticeship. However, teachers do have knowledge and expertise in pedagogy, managing classrooms, and creating climates in which ideas can be freely expressed and explored (Frederiksen et al., 1997; Peterson, 1988; Shulman, 1987). Our approach, therefore, is to build on this expertise by carefully scaffolding the creation of a classroom research community through providing detailed and explicit lesson plans, computer simulations and activities, and student research materials. We then rely on the same process of implementation coupled with reflection to allow both teachers and students to develop a mature understanding of scientific inquiry.

THE THINKERTOOLS INQUIRY CURRICULUM

The Target Expertise

Our curriculum attempts to transform a science classroom into a research community whose objective is to "discover" the laws of force and motion. The research is guided by a scientific inquiry cycle in which the goal is to develop an explicit, written-down, conceptual model of how forces affect motion. It is also guided by a self-assessment process in which students continually monitor and reflect on their progress.

Inquiry Expertise

Our view of inquiry expertise is for students to learn to do the following:

- Formulate a well-formed, investigatable research question whose pursuit will advance their understanding of a topic they are curious about (QUESTION).
- Generate alternative, competing hypotheses and predictions about what might happen with respect to that question and why it might happen (PREDICT).
- Design and carry out experiments using both the real world and computer simulations in order to determine what actually happens (EXPERIMENT).
- Analyze their data to construct an explicit conceptual model that includes scientific laws that would predict and explain what they found (MODEL).
- Apply their model to different situations in order to investigate its utility as well as its limitations (APPLY)—the limitations raise new research questions and the cycle repeats (QUESTION).

The idea is for the students to go through this Inquiry Cycle for each research topic in the curriculum and to develop, by the end of the curriculum, the skills to do independent research on topics of their own choosing. To facilitate this development, students need to acquire the metacognitive expertise of being able to monitor and reflect on their research. They learn to use criteria, such as “Reasoning Carefully,” to evaluate their progress at each step of the Inquiry Cycle and after each completion of the cycle. As they continuously engage in this Reflective Self-Assessment Process, their knowledge of inquiry should develop and improve.

**Physics Expertise**

Our view of physics expertise is centered around enabling students to construct and utilize causal models at an intermediate level of abstraction (B. White 1989, 1993a). These models increase in their complexity and sophistication as the students progress through the curriculum. For example, in Module 1, students develop a conceptual model for predicting what will happen in a simplified, idealized, one-dimensional world with no friction or gravity. Their model enables them to step through time and analyze what forces are acting on an object and then, based on those forces, to determine the object’s velocity and position. They achieve this by developing explicit laws and representations to calculate and encode what is happening. For example, the following are versions of Newton’s first two laws of motion:

- **Basic Principle**: An object’s speed stays the same except when you apply an impulse, and, whenever you apply an impulse, it changes speed.
- **Prediction Law**: If the impulse is in the same direction that the object is moving, it adds one to its speed (+1); in the opposite direction, it subtracts one from its speed (−1).

The basic principle tells you what events to pay attention to, and the prediction law tells you how to calculate the effects of those events. In conjunction with such laws, students create and use representations to calculate and show changes in velocity. For instance, in Figure 2, the student was asked to predict the effects of a sequence of impulses on an object’s motion: The arrows show when an impulse was applied, the “datacross” helps to calculate and show the effect that the impulse had on the object’s velocity, and the “dotprints” show the object’s position and velocity as time passes.

---

2Impulses are forces that act for a specified—usually short—amount of time like a kick or a hit.

3A datacross is a speedometer-like representation that indicates the horizontal and vertical components of an object’s velocity.

4Dotprints are a “trail” of small dots that an object leaves behind as it moves (like footprints). Because they are put down at regular time intervals, such as once per second, they show by their position and relative separation the history of an object’s velocity.
FIGURE 2 A model-based prediction of the effects of a sequence of impulses on an object's motion.

The target physics expertise includes understanding and using these semiabstract representations and causal laws and embodying them in a conceptual model that lets students step through time to envision, predict, and explain what is happening. Our objective is for students to understand the form and properties of such scientific laws and models, the inquiry process needed for creating them, and the utility of such models for predicting, controlling, and explaining real-world behavior.

An Overview of the Instructional Approach

To enable students to achieve this expertise, we developed a constructivist approach that could be characterized as learning metacognitive knowledge and skills through a process of scaffolded inquiry, reflection, and generalization. This approach is centered around the Inquiry Cycle, which is used to guide the students’ research. The curricular activities focus on enabling students to develop the expertise needed to carry out the steps in the Inquiry Cycle as they conduct their research, as well as to monitor and reflect on their progress.

1. Scaffolded Inquiry: We designed scaffolded activities and environments to enable students to learn about inquiry as they engage in authentic scientific research. The scaffolded activities help them learn about the characteristics of scientific laws and models, the processes of experimental design and data analysis, and the nature of scientific argument and proof. The scaffolded environments, which include computer simulations (which allow students to interact with models of force and motion) and analytic tools (for analyzing the results of their computer and real-world experiments), make the inquiry process as easy and productive as possible at each stage in learning. These activities and environments enable students to carry out a sequence of activities that correspond to the steps in the Inquiry Cycle. Initially, the meaning and purpose of the steps in the Inquiry Cycle may be only partially understood by students.

2. Reflective Assessment: In conjunction with the scaffolded inquiry, students are introduced to a reflective process in which they evaluate their own and each other’s research. This process employs a carefully chosen set of criteria that characterize expert scientific inquiry (such as “Understanding the Inquiry Process,”
"Being Systematic," and "Reasoning Carefully") to enable students to see the intellectual purpose and properties of the inquiry steps and their sequencing. By reflecting on the attributes of each activity and its function in constructing scientific theories, students grow to understand the nature of inquiry and the habits of thought that are involved.

3. Generalized Inquiry and Reflection: The Inquiry Cycle is repeated as the class addresses new research questions. Each time the cycle is repeated, some of the scaffolding is removed so that eventually the students are conducting independent inquiry on questions of their own choosing. These repetitions of the Inquiry Cycle in conjunction with reflection help students to refine their inquiry processes. Carrying out these processes in new research contexts also enables students to learn how to generalize the inquiry and reflection processes so that they can apply them to new learning situations.

The ThinkerTools Software

One of the primary conceptual tools that the students have to work with is the ThinkerTools software that we developed for the Macintosh computer. This software enables students to interact with Newtonian models of force and motion (see Figure 3). The software also lets students create their own models and experiments. They use drawing tools to create and run computer simulations. Objects (such as

FIGURE 3 The ThinkerTools software provides modeling and inquiry tools for creating and experimenting with models of force and motion.
the large circle shown in Figure 3) and barriers can be placed on the screen. (The objects are introduced to students as generic objects, simply called dots; they are the pictorial equivalent of a variable so that students can map them on to different objects such as spaceships or billiard balls.) Students can define and change the properties of any object, such as its mass, elasticity (e.g., bouncy or fragile), and velocity. Students can then apply impulses to the object to change its velocity using the keyboard or a joystick as in a video game. Students can thus create and experiment with a “dot-impulse model,” and they can discover, for example, that when one applies an impulse in the same direction that the dot is moving, it increases the dot’s velocity by one unit of speed. In this way, they can use simulations to discover the laws of physics and their implications.

The ThinkerTools software enables students to create experimental situations that are difficult or impossible to create in the real world. For example, they can turn friction and gravity on and off and can select different friction laws (i.e., sliding friction or gas–fluid friction). They can also vary the amount of friction or gravity to see what happens. Such experimental manipulations in which students dramatically alter the parameters of the simulation allow students to use inquiry strategies, such as “look at extreme cases,” which are hard to utilize in real-world inquiry. This type of inquiry enables students to see more readily the behavioral implications of the laws of physics and to discover the underlying principles.

Another advantage of having students experiment with such simulations is that the software includes measurement tools that allow students to easily make accurate observations of distances, times, and velocities. These observations would often be very difficult to make in the corresponding real-world experiment. The software also includes graphical representations of variables. For example, as the dot moves, it leaves behind dotprints, which show how far it moved in each second. There is also the datacross, which shows its X and Y velocity components. And, students can have the software keep a table or graph to record, for example, the velocity of the dot. In addition, there are analytic tools such as “stepping through time,” which allows students to pause the simulation and to proceed time step by time step so that they can better see and analyze what is happening to the motion of the dot. In this mode, the simulation runs for a small amount of time, leaves one dotprint on the screen, and then pauses again. The students have control over whether the simulation remains paused, proceeds to the next time step, or returns to continuous mode. These analytic tools and graphical representations help students determine the underlying laws of motion. They can also be incorporated within the students’ conceptual model to represent and reason about what might happen in successive time steps. In this way, such dynamic interactive simulations can provide a transition from students’ intuitive ways of reasoning about the world to the more abstract formal methods that scientists use for representing and reasoning about the behavior of a system (B. White, 1993a).
A further advantage of computer simulations is that they help students to understand the nature of both scientific and conceptual models. After all, the computer is not the real world; it can only simulate real-world behavior by stepping through time and using rules to determine how any forces that are acting (like friction or gravity) will change the dot’s velocity on that time step. Thus, the computer is actually using a conceptual model to predict behavior, just as the students will use the conceptual model they construct to predict behavior. In working with the computer, the students’ task is to design experiments that will help them discover the laws that are used by the simulation. This is more straightforward than the corresponding real-world inquiry task. After all, objects in the real world are not driven by laws; rather, the laws simply characterize their behavior.

One example of a modeling activity, carried out early in the curriculum, asks students to explain how their computer and real-world experiments could lead to different conclusions. They might say, for instance, that “the computer simulation does not have friction, which is affecting our real-world experiments.” Alternatively, they might say that “the real world does not behave perfectly and does not follow rules.” Working with a computer simulation can thus help students to develop metaconceptual knowledge about what scientific conceptual models are and how laws can be used to predict and control behavior. It can also enable them to appreciate the utility of creating computer simulations that embody scientific laws and idealized abstractions of real-world behavior and, then, of using such simulations to do experiments in order to see the implications of a particular theory.

The Instructional Implementation of the Inquiry Process

We now briefly describe how the teaching and learning of inquiry is implemented within the curriculum. In order to provide a better sense of how the curriculum proceeds, we will organize this description around the five steps of the Inquiry Cycle.

**Question**

Formulating a good research question that allows you to do simple experiments to advance your theory is very hard—graduate students frequently find this the most difficult aspect of doing a doctoral thesis. So, in our curriculum, this process is heavily scaffolded for the students. The curriculum employs the research strategy of starting with simplified idealized worlds, such as a world of one-dimensional motion in which there is no friction or gravity, and then of gradually adding complexities.
In the Question Phase of the Inquiry Cycle, this strategy is made explicit to the students. The teacher starts the curriculum by having the students toss a ball around the room, while at the same time observing and listing all of the factors that may be involved in determining its motion (how it is thrown, gravity, air resistance, etc.). They come to realize that this apparently simple situation is actually very complicated. As a research strategy, the teacher suggests the need to simplify the situation, and this discussion leads to the idea of looking at simpler cases, such as that of one-dimensional motion where there is no friction or gravity (e.g., a ball moving through outer space). This activity thus helps students appreciate the need to start, in Module 1, with trying to understand simple one-dimensional motion and, then, in successive modules, to study complicating factors such as friction, mass, two-dimensional motion, and gravity. An outline of the curriculum is presented below.

The ThinkerTools Curriculum

Module 1. One-Dimensional Motion
Module 2. Friction
Module 3. Mass Project (the Common Inquiry Project)
Module 4. Two-Dimensional Motion
Module 5. Gravity
Module 6. Trajectories
Module 7. Final Project (the Chosen Inquiry Project)

Through this research strategy, students gradually create a conceptual model that they can use to analyze complex projectile motion like that of the ball being thrown around the room. At the end of the curriculum, they are presented with a variety of possible research topics to pursue, and they carry out research projects on topics of their own choosing. Examples of research topics include “Investigating Circular Motion,” “Analyzing Collisions,” and “Exploring Projectile Motion in a Different World” (such as one with a higher gravity and denser atmosphere).

**Predict**

Once a general research question has been identified, such as “what happens in a one-dimensional world with no gravity or friction,” the class moves on to the next phase in the Inquiry Cycle. In this phase, the teacher asks students to make predictions about what they think might happen in some simple real-world situations that are related to the research question. In other words, they are asked to engage in “thought experiments” (Horowitz & Massey, 1991; T. Kuhn, 1964, 1970; Sorensen, 1992). For example, given the general question about a one-dimensional world with no gravity or friction, the teacher might ask them to think more
specifically about what would happen when you hit a ball on a frictionless surface. The students individually record their predictions in their own research books. They are also asked to record an alternative hypothesis that they or someone else might have, as shown in the following:

*Predictive Question:* Imagine that a ball is stopped on a frictionless surface. Suppose that you hit the ball. Then, right after the hit, you hit the ball again in the opposite direction with the same size hit.

Would the hit in the opposite direction change the velocity of the ball? If so, describe how it would change and explain why.

Think of a different hypothesis that you or someone else might have about how the ball's velocity would change. Describe it and explain why someone might believe it.

The teacher then asks different students to present their ideas about what they think might happen and why it might happen. For instance, in response to the previous question, one student might say, "The ball will turn around and go in the opposite direction because that is the way you hit it"; another might say, "The ball will stop because the two hits cancel out"; and yet another might say, "The ball will slow down because the hit in the opposite direction took away some of its speed."

By having the students present and explain their predictions, the teacher gets the class to generate a set of alternative theories about what might happen. The predictive questions thus help students to develop more focused research questions to investigate and to design experiments that differentiate among the possible theories generated by the class.

*Experiment*

Once the class has generated some different predictions and theories about what might happen, they then go on to the Experiment Phase of the Inquiry Cycle. In this phase, each student works with a research partner to try to determine what actually does happen. Their purpose is to find out through appropriately designed experiments which of the competing hypotheses, if any, is accurate. Here, there are a variety of different experimental media that the students use to conduct their research. These include computer simulations, the real world itself, and videos of the real world.

*Computer simulations.* Using the ThinkerTools software, students can do a variety of experiments and game-like activities in order to discover the laws that
FIGURE 4  The goal of this game is to make the dot hit the target with a specified velocity. In this case, the dot needs to be moving to the right with three units of speed when it hits the target.

govern the simulated world. One example is the activity shown in Figure 4 in which students have to control the motion of the dot in a simple one-dimensional world without friction and gravity. They control the dot’s motion by applying impulses to it. If the dot hits the target X while traveling at the specified speed, the target “catches” the dot. If not, the dot crashes into the end barrier and explodes. This game helps students to discover the additive effects that impulses have on the dot’s velocity and works to counteract the common misconception that motion requires a force (“Unless you keep applying impulses, the dot will slow down”) and to replace it with the Newtonian conception that motion does not require a force and forces cause accelerations (“An object’s speed stays the same except when you apply an impulse, and whenever you apply an impulse, it changes velocity”).

The students also work with the Newtonian computer simulations in a more experimental and less game-like fashion. For example, in Module 2, students attempt to determine a law for the effects of sliding friction. To facilitate this, they work with a computer simulation in which they can change the initial velocity of a dot and can also change the amount of friction. By manipulating these variables and analyzing what happens (using dotprints, tables, and graphs), they determine, for instance, whether the effects of sliding friction are dependent on the velocity of the dot.
**The real world.** Another medium for experimentation is the real world. Early on in the curriculum, in Module 1, students are given an experiment to do, like the one shown in Figure 5 called the Double Bonk Experiment. In this experiment, students give a billiard ball a single bonk versus a double bonk to see what happens to its velocity. This activity provides them with an example of a real-world experiment. In Module 2, they have to criticize a flawed experiment (shown at the end of Appendix A) as well as carry out an example of an unflawed experiment. From Module 3 onward, they have to design their own real-world experiments (see Appendix B). To guide them in planning and analyzing their experiments, they are presented with an outline such as the following:

**Research Guide**

For each experiment, you need to do the following:

1. Create a plan with:
   - A sketch showing how you will set up the equipment.
   - A description of what you will do and how you will measure the velocity of the ball.

2. Do your experiment.
   - Record your data in a clear and organized way.
   - Record any problems you had in doing your experiment.

3. Analyze your data and present your conclusions.
   - State any laws you discovered that predict and describe what happens.
   - Give an explanation for why this happens.
   - Explain how your results agree or disagree with what you predicted would happen when you stated your hypotheses.

---

**The Double-Bonk Experiment**

![Diagram of the Double Bonk Experiment](image)

**Data Table**

<table>
<thead>
<tr>
<th>Trial</th>
<th>Distance (cm)</th>
<th>Time (sec)</th>
<th>Velocity (V=D/T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 5** An illustration of the Double Bonk Experiment along with the data table that students use to record their experimental results.
Through this technique of progressively removing scaffolding, the process of experimental design is introduced to students, and they learn how to do their own experiments.

*Video.* The real world is useful for learning about experimental design and measurement error, but there are limits to what students can discover from such experiments (Gunstone, 1990). For example, the real world is hard to control, and so it is difficult for students to do their experiments and get accurate results. They often produce data that are messy and uninterpretable. The real world also does not enable them to step through time to see how an object’s velocity changes as time passes.

To overcome these problems, the students need better experimental technique and a better recording medium such as videotape. Videotape allows them to see how velocity changes with time: Students can place a transparency on the video monitor and can step through the videotape and draw dotprints on the transparency at fixed time intervals to see how the motion of the object on the tape is changing (see Figure 6). Students can then analyze the dotprints from the video in addition to the data from their own real-world experiments. This gives them a cleaner and richer set of data. Ideally, students would use videotapes of their own experiments for this purpose, but we did not have the resources to make this possible. Instead, we provide them with videotapes of prototypical motions, like an object sliding across a surface or being hit off a cliff. These videos and their resulting dotprints enable students to carry out more in-depth analyses of how objects behave in the real world and to better compare real-world behavior with how objects behave in the computer simulation.

*Model*

Once the students have finished running their experiments, they progress to the Model Phase of the Inquiry Cycle. In this phase, students have to analyze their data
and characterize what they found in the form of a conceptual model (which incorporates explicit laws and representations and a method for envisioning what happens). The key components of this conceptual model, which the students make explicit, are "prediction laws." In the early stages of the curriculum, in Module 1, this modeling process is scaffolded. For example, students are presented with alternative possible laws to evaluate (as shown in Appendix A). In this way, they are presented with examples of "good" and "bad" laws, and of different types of laws, before they have to construct their own.

The laws that students have to evaluate in Module 1 were carefully designed to provoke discussions of the correctness and precision of a rule's predictions, the range of situations to which it applies, and its simplicity and memorability. This set of laws also illustrates the different types of laws that are needed in a conceptual model. For example, consider one of the qualitative laws: "Whenever you apply an impulse to the dot, it changes speed" (which is a version of Newton's first law). Also, consider one of the quantitative laws: "If you apply an impulse in the same direction that the dot is moving, it adds one to its speed; in the opposite direction, it subtracts one" (which is a version of Newton's second law, $F = ma$). Typically, some students will argue that you need both of these laws. In their conceptual model, the first law is a necessary part of their model's control structure that tells them what events to pay attention to, and the second law is the type of law needed to make precise predictions about what will happen if those events occur. Newton felt that he needed both laws even though many physicists argue that they are redundant (Steinberg, Brown, & Clement, 1990). From the point of view of the conceptual model the students create, they are not at all redundant—both kinds of laws are needed.

From Module 2 onward, students work with their research partner to construct such laws for themselves. For example, in Module 2 (friction), the students' research book asks the following:

Based on the results of your experiments, create a "friction prediction" rule. What is your theory about why this happens?

Once all of the pairs have completed their research and formulated a law, the class gets together to try to reach a consensus about which laws best account for their results and are the most accurate and useful. As part of this process, they have to critique each other's laws and attempt to prove them wrong. For example, in Module 4, in which the students investigate two-dimensional motion, they engage in the following activities:

*Creating your own laws*: Work with your partner to invent two different laws that predict what will happen to the dot in situations like those of the Predictive Questions in which you apply impulses at right angles.
• Make one law that is correct and one that is incorrect.

_Evaluating another group’s laws:_ Trade your laws with those created by another group, so that you can evaluate each other’s laws. Work with your partner to determine which of the other group’s laws is incorrect and which is correct. After each law, give your evaluation of that law:

• For the law that you think is incorrect, prove that it is incorrect.
• For the law that you think is correct, say how useful it is and explain your reasons.

After the class has finished these law creation and evaluation activities, all of the laws are put on the board. The class then attempts to reach agreement about which laws are incorrect. If there are disagreements, the students who believe a law to be incorrect are asked to present their proofs to the class. Then, for the correct laws, the class attempts to reach agreement about which are the most useful. Again, if there are disagreements, students are asked to present arguments in favor of their choice(s) for the most useful law(s). Activities such as these foster discussions about the nature of scientific models and proof. They also play an important role in creating a classroom research community.

_Apply_

Once the students have created their explicit conceptual model in the form of representations and laws for predicting behavior, they then go on to the Apply Phase of the Inquiry Cycle. In this phase, they try to apply their conceptual model to different real-world situations. As part of this process, they investigate its utility for predicting and explaining what would happen, and they also investigate the limits of their model.

For instance, in the Experiment Phase of Module 4 (two-dimensional motion), students construct laws for predicting the change in motion that occurs when an orthogonal impulse is applied, and the class then gets together to decide which of these is the most useful (an example of such a law is illustrated in Figure 7). Then, in the Apply Phase of the Inquiry Cycle, they try to apply the law to different real-world situations. For instance, students might illustrate how it could be used in playing soccer. (One student explained, for example, how the law shown in Figure 7 could be used to determine which way to run to intercept a fast or slow moving ball that is being kicked by another player; this type of prediction is illustrated in Figure 7.)

The teacher then motivates the next research topic, in this case “gravity,” by asking the class if their law would make accurate predictions if the soccer game were being played on a steep hill or if the ball were kicked up in the air. Hopefully,
When you give a moving dot a sideways impulse, it turns; the slower it is moving when you apply the impulse, the more it turns.

FIGURE 7 An example of a student’s two-dimensional prediction rule for a world with no friction or gravity.

someone answers that the predictions would be incorrect because their model does not take gravity into account. So, investigating the limits of their model raises a new research question. This brings the class back to the beginning of the Inquiry Cycle and to investigating the next research question in the curriculum, which in this case is, “What are the laws for gravity?” The class then proceeds, following the Inquiry Cycle, to determine laws for gravity, to incorporate them into their conceptual model, and to use them to analyze situations like playing soccer on a hill or kicking a ball up in the air.

Cycling Toward Independent Inquiry

The Inquiry Cycle is thus repeated within each of the seven modules of the curriculum. The physics the students are dealing with increases in complexity as the curriculum progresses and so does the inquiry. In the early stages of the curriculum, the inquiry process is heavily scaffolded. For example, in Module 1, students are given experiments to do and are presented with alternative possible laws to evaluate. In this way, they see examples of experiments and laws before they have to create their own. In Module 2, students are given experiments to do but have to construct the laws for themselves. Then, in Module 3, they design their own experiments and construct their own laws to characterize their findings (see
Appendix B). By the end of the curriculum, the students are carrying out independent inquiry on a topic of their own choosing as well as presenting research reports to the rest of the class.

Reflective Assessment and the Learning of Inquiry

In addition to the Inquiry Cycle, which guides the students' research and helps them to understand what the research process is all about, we also developed a set of criteria for characterizing good scientific research. These are presented in Figure 8. They include goal-oriented criteria such as "Understanding the Science" and "Understanding the Processes of Inquiry," process-oriented criteria such as "Being Systematic" and "Reasoning Carefully," and socially-oriented criteria such as "Communicating Well" and "Teamwork." These characterizations of good work are used not only by the teachers in judging the students' research projects but also by the students themselves.

At the beginning of the curriculum, the criteria are introduced and explained to the students as the "Criteria for Judging Research" (see Figure 8). Then, at the end of each phase in the Inquiry Cycle, the students monitor their progress by evaluating their work on the two most relevant of these criteria (see Figure 9). And then, at the end of each module, they reflect on their work by evaluating themselves on all of them (see the end of Appendix B). Similarly, when they present their research projects to the class in Modules 3 and 7 of the curriculum, the students evaluate not only their own research projects but also each other's. They give each other feedback both verbally and in writing. These research criteria are thus used as a way of helping to introduce the students to the characteristics of good research, and to monitoring and reflecting on their inquiry process.

Appendix B presents an example of a student's research project along with her self-assessment of her work. In what follows, we present sample excerpts from a class's reflective-assessment discussion in which students provide feedback to classmates who have just given an oral presentation of their research project. Excerpts are chosen so as to provide a sample reflection for each of the scoring criteria. (Pseudonyms are used throughout, and the transcript has been lightly edited to improve its readability.)

Teacher: OK, now what we are going to do is give them some feedback. So, who can comment on a scale from 1 to 5 on their "understanding the science"? Mark.

Mark: I gave them a 4 because, although they were sort of shy in their presenting, it seems like they understand it pretty well. Presenting it seemed the hard part, but they look like they understand their science.
Criteria for Judging Research

UNDERSTANDING

**Understanding the Science.** Students show that they understand the science developed in the curriculum and can apply it in solving problems, in predicting and explaining real-world phenomena, and in carrying out inquiry projects.

**Understanding the Processes of Inquiry.** Students can talk about what approach they or others have taken in exploring a research topic. For instance, they can explain what types of scientific models and inquiry processes have been used in carrying out investigations and in reaching conclusions.

**Making Connections.** Students see the big picture and have a clear overview of their work, its purposes, and how it relates to other ideas or situations. They relate new information, ideas, and experimental results to what they already know.

PERFORMANCE: DOING SCIENCE

**Being Inventive.** Students are creative and examine many possibilities in their work. They show originality and inventiveness in thinking of problems to investigate, in coming up with hypotheses, in designing experiments, in creating new laws or models, and in applying their models to new situations.

**Being Systematic.** Students are careful, organized, and logical in planning and carrying out their work. When problems come up, they are thoughtful in examining their progress and in deciding whether to alter their approach or strategy.

**Using the Tools of Science.** Students use the tools and representations of science appropriately. The tools they choose to use (or create) may include such things as lab equipment, measuring instruments, diagrams, graphs, charts, calculators, and computers.

**Reasoning Carefully.** Students can reason appropriately and carefully using scientific concepts and models. For instance, they can argue whether or not a prediction or law that they or someone else has suggested fits with a scientific model. They can also show how experimental observations support or refute a model.

SOCIAL CONTEXT OF WORK

**Writing and Communicating Well.** Students clearly express their ideas to each other or to an audience through writing, diagrams, and speaking. Their communication is clear enough to allow others to understand their work and reproduce their research.

**Teamwork.** Students work together as a team to make progress. Students respect each others' contributions and support each others' learning. Students divide their work fairly and make sure that everyone has an important part.

FIGURE 8 The criteria for judging research that students use in the Reflective-Assessment Process.
Now you will evaluate the work you just did.

**Being Systematic**

**Being Systematic.** Students are careful, organized, and logical in planning and carrying out their work. When problems come up, they are thoughtful in examining their progress and deciding whether to alter their approach or strategy.

Circle the score that you think your work deserves

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not adequate</td>
<td>adequate</td>
<td>exceptional</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Explain how your experimental work and analyses justify the score you have given yourself.

Writing and Communicating Well

**Writing and Communicating Well.** Students clearly express their ideas to each other or to an audience through writing, diagrams, and speaking. Their communication is clear enough to allow others to understand their work and reproduce their research.

Circle the score that you think your work deserves

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not adequate</td>
<td>adequate</td>
<td>exceptional</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Explain how your experimental write-ups justify the score you have given yourself.

FIGURE 9 An example of a self-assessment page found in the students’ research books. This sample page is located at the end of the Experiment Phase of the Inquiry Cycle.

Teacher: OK, what about their “understanding the process of inquiry”? In terms of their following the steps within the inquiry cycle, on a scale from 1 to 5, how would you score them? Vanessa.

Vanessa: I think I would give them a 5 because they followed everything. First, they figured out what they wanted to inquire, and then they made hypotheses, and then they figured out what kind of experiment to do, and then they tried the experiment, and then they figured out what the answer really was and that Jamal’s hypothesis was correct.
Teacher: Great. OK, what about their “making connections”? On a scale from 1 to 5, who would like to talk about that? Brandon?

Brandon: I think a 5. I think they really understood how it connected with the real world, like with football and all. I really liked what they said about the football players: if one is heavier than the other, the big one will just run over the little one, but if they’re the same mass, they might bounce back.

Teacher: All right, in terms of their performance, “being inventive.” Justin?

Justin: Being inventive. I gave them a 5 because they had completely different experiments than almost everyone else’s I’ve seen. So, being inventive, they definitely were very inventive in their experimentation.

Teacher: OK, what about “being systematic”? Emily.

Emily: I think I would give them a 4 because it sort of looked like they skipped some parts of what they were supposed to do.

Teacher: OK, Carla [one of the presenters], how would you evaluate yourself?

Carla: I gave myself a 4 because I was organized in my work most of the time. And, we did all the steps that we were supposed to do for our project. And, we summarized them in our presentation.

Teacher: Good. What about, where am I? Oh, “using the tools of science.” Nick.

Nick: Well, I know it would be hard, but they weren’t able to get their real-world experiment to work. So, I’d give them a 3. Otherwise, if they had gotten that to work better, then definitely it would have been a 5.

Teacher: OK, Jamal [one of the presenters], how would you evaluate yourself?

Jamal: OK, I gave myself a 4 because we always used everything perfectly except when we couldn’t get to the computers.

Teacher: OK, good. What about “reasoning carefully”? Jamal, how would you evaluate yourself on that?

Jamal: I gave myself a 5, because I had to compute the dotprints between the experiments we did on mass. So, I had to compute everything. And, I double-checked all of my work.

Teacher: Great. OK, in terms of the social context of work, “writing and communicating well.” Carla, how did you score yourself in that area?

Carla: I gave myself a 4, because I always told Jamal what I thought was good or what I thought was bad, and if we should keep this part of our experiment or not. We would debate on it and finally come up with an answer.

Teacher: What about “teamwork”? Does anyone want to rate that? Teamwork.

Nisha: I don’t know if I can say because I didn’t see them work. [laughter]
Teacher: That's fine. That's fair. You are being honest. Julia?

Julia: I gave them a 5 because they both talked in the presentation, and they worked together very well, and they looked out for each other.

There are various arguments for why incorporating such reflective-assessment processes into the curriculum should be effective. One is the “transparent assessment” argument put forward by Frederiksen and Collins (1989; see also Frederiksen, 1994), who argue that introducing students to the criteria by which their work will be evaluated enables students to better understand the characteristics of good performance. In addition, there is the argument about the importance of metacognition put forward by researchers (e.g., Baird, Fensham, Gunstone, & White, 1991; A. Brown et al., 1983; A. Brown & Campione, 1996; Bruer, 1993; Collins, et al., 1989; Miller, 1991; Nickerson et al., 1985; Reeve & Brown, 1985; Resnick, 1987; Scardamalia, Bereiter, & Lamon, 1994; Schauble & Glaser, 1990; Schauble, Raghavan, & Glaser, 1993; Schoenfeld, 1987; Schön, 1983, 1987; Towler & Broadfoot, 1992) who maintain that monitoring and reflecting on the process and products of one’s own learning is crucial to successful learning as well as to “learning how to learn.” Research on good versus poor learners shows that many students, particularly lower achieving students, have inadequate metacognitive processes, and their learning suffers accordingly (Campione, 1987; Chi, Bassock, Lewis, Reimann, & Glaser, 1989). Thus, if you introduce and support such processes in the curriculum, the students’ learning and inquiry should be enhanced.

Instructional trials of the ThinkerTools curriculum in urban classrooms (which include many lower achieving students) provided an ideal opportunity to test these hypotheses concerning the utility of such a reflective peer and self-assessment process. For each of the participating teachers, one half of his or her classes engaged in the Reflective-Assessment Process and the other half did not. In other words, each teacher taught control classes as well as reflective-assessment classes. The “control” classes did the same ThinkerTools curriculum but commented, at the end of each module, on what they did and did not like about that module, as opposed to the “experimental” classes, who reflected, as described earlier, on their own research process and products.

INSTRUCTIONAL TRIALS: METHODS AND PARTICIPANTS

Having completed an overview of the instructional approach, the ThinkerTools curricular materials, and how they are used to teach science, we now turn to describing the instructional trials of the curriculum. The curriculum was implemented in the spring of 1994 by three teachers in 12 urban classes in two schools. In these schools, students have a 45-min science period each school day. On average, the ThinkerTools Inquiry Curriculum took 10.5 weeks to complete.
Teachers

The participating teachers included two teachers who had worked with us for 1 month to help develop the materials and who had subsequently piloted the materials in their classes and made recommendations for their revision. Also included was a third teacher who had had no contact with us prior to implementing the curriculum. Teacher 1 had studied physics in high school and college. Teachers 2 and 3 had no prior courses in physics. Because Teacher 3 was not involved in the curriculum design process, she provides the best test of whether these materials are usable by a teacher with no physics training.

Classes and Students

Teacher 1 implemented ThinkerTools in 3 seventh-grade classes and 2 ninth-grade classes, Teacher 2 in 3 eighth-grade classes, and Teacher 3 in 4 seventh-grade classes. Teachers 2 and 3 were in the same school.

All three teachers are teaching in urban schools. Their class sizes averaged almost 30 students. Two thirds of the students were minority students. In one of the schools (that of Teachers 2 and 3), as many as 25% of the students were from families who receive Aid for Dependent Children. The distribution of Comprehensive Test of Basic Skills (CTBS) percentile scores for the 12 classes is nearly flat over the range from the 1st to the 99th percentile, indicating that the students' achievement levels approximate those of a national sample. However, the distribution is a little top heavy, with a median percentile score of 60. This wide distribution is ideal for research purposes but is challenging for teachers. These teachers were thus working with a very diverse population of students with widely varying prior academic achievement.

There were no significant differences in students' average CTBS scores for the classes that were randomly assigned to the different treatments (reflective assessment vs. control), for the three different teachers, or for the different grade levels (seventh, eighth, and ninth). Thus, the classes were all comparable with regard to achievement test scores. Also, the average CTBS scores for male and female students were comparable.

Teacher Supports

The primary support we gave the teachers was to provide them with all of the curricular and instructional materials required: computers, lab equipment, students'
research books (which include all of the scaffolded inquiry activities), assessments, and teacher’s guides (which attempt to scaffold the teaching of inquiry for the teachers). Research assistants were sometimes present in the classrooms for purposes of videotaping, but they provided no assistance with classroom instruction. The teachers often came to our weekly research group meetings (usually attended by the two project principle investigators, four research assistants, and the programmer) for about 1 hr to report on how things were going in their classrooms and to discuss how the curriculum could be improved. In addition, when they came for these meetings, they also occasionally met as a subgroup (consisting of only the three teachers) so that they could look at videotapes of their classes and discuss them. Through these meetings with us and each other, the teachers had an opportunity to reflect on their teaching. Aside from providing them with the scaffolded curricular materials, inviting them to attend our weekly meetings, and giving them videotapes of their classes to discuss with each other, no other supports were provided for the teachers by our research team.

Data Collection and Analyses

Data collected in these instructional trials include pretests and posttests measuring students’ inquiry skills, physics knowledge, and attitudes about learning science. We also collected their research books and projects. Furthermore, we made videotapes of a sample of the classes (taken evenly across the different classes) and conducted interviews with some of the students at the end of the curriculum.

The data for students’ understanding of inquiry and physics were analyzed using analyses of variance (ANOVA) that included four between-subjects factors:

- Treatment (reflective assessment and control)
- Gender (male and female)
- CTBS achievement level (low [≤ 60] and high [> 60])
- Grade level (7 and 8/9)

There were two teachers within each level of the grade factor, one of whom taught both seventh- and ninth-grade classes. When a test was given as both a pretest and a posttest, the time of test (pre and post) was included as an additional, repeated measures factor. One-tailed tests were carried out in our analyses of the effects of reflective assessment because our hypothesis is that such a metacognitive process

---

*With regard to achievement levels, we used the percentile rank of the combined CTBS verbal and quantitative scores to classify students. We used two levels of CTBS: low (0–60) and high (61–100). These cut scores for the two CTBS levels were chosen to divide our sample into halves as evenly as possible with regard to all of the different factors (i.e., treatment, grade level, etc.).*
will be beneficial, particularly for students from more disadvantaged backgrounds. To test the differential effects of reflective assessment for different levels of student achievement, we carried out planned comparisons of reflective-assessment and control treatments for each achievement level. Although we refer to CTBS scores as measures of students’ prior achievement, we regard the CTBS scores as a proxy for students’ more general level of educational advantage or disadvantage.

RESULTS: INQUIRY EXPERTISE

We turn now to presenting results related to the students’ development of scientific inquiry skills. In analyzing the inquiry tests and research project data, we will address the following general questions:

1. How well did students develop skill at scientific inquiry?
2. How did the Reflective-Assessment Process affect the learning and doing of inquiry?
3. How did the students’ grade level or achievement level affect their inquiry performance?

Research Projects

Our first analyses are based on evaluations of students’ inquiry projects, which included the Mass Project (in which all students addressed the same research question) and the Final Project (in which students chose their research question from among nine different topics). The Mass Projects were completed about halfway through the curriculum, and the Final Projects were completed at the end of the curriculum. For each project, an ANOVA was carried out on project scores assigned by the teachers. Data were analyzed for students who completed their research project, for whom the projects were scored, and for whom we also have CTBS scores.

Scoring. The students’ reports for their Mass Projects (Module 3) and their Final Projects (Module 7) were scored holistically on a 5-point scale ranging from 1 (not adequate) to 5 (exceptional) for their overall quality, as well as for each of the criteria for judging research shown in Figure 8. Scoring was done by the teachers after the close of the school year as part of a scoring study to investigate the reliability and validity of the scoring system (Frederiksen & White, 1997b). In order to avoid the possibility of biasing the teachers, they were not told that the scores from this scoring reliability study would also be used to assess the impact of the Reflective-Assessment Process. In addition, they were not told which classes the students were
in—of course a teacher might recognize her own students, but the high reliability of
the scores across teachers indicates that there was no detectable bias.

The research reports were scored for eight of the classes (one treatment and one
control class for each teacher at each grade level) in a randomized order. A sample
of 48 reports (stratified to represent all teachers and conditions equally) was scored
by all three teachers. For comparison purposes, half of these were scored at the
beginning and half at the end of the scoring study. After the initial set was scored
by all three teachers (and the teachers had become consistent in their scoring), the
bulk of the projects were scored by one or two teachers. Then, at the end of the
scoring study, a final set was scored by all three teachers to enable us to determine
whether there were changes in reliability with practice. In all, 44% were scored by
one teacher, 11% by two teachers, and 45% by three teachers. When projects were
scored by more than one teacher, the teachers first scored them individually and
then discussed their scores and recorded their moderated scores. The scores
included in this analysis were the average moderated scores for projects that were
scored by more than one teacher and individual scores for projects scored by a single
teacher. In all, 75% of the students’ projects in the eight classes were scored.

Module 3: Common Research Projects

Probably our best single measure of inquiry is the Common Research Project
that students carried out in Module 3, also known as the Mass Project. In this project,
which was completed approximately halfway through the curriculum, students
investigated the relationship between the mass of an object and the effect that an
impulse has on its motion. Appendix B presents the instructions for this project and
an example of one student’s project as well as her reflective self-assessment.

Scoring. The Mass Projects were scored holistically on a 5-point scale for
overall quality, as well as for each of the criteria for judging research shown in
Figure 8. The projects were scored in a randomized order by one to three teachers,
as described earlier. The distribution of scores for overall quality is given in Figure
10 to show how frequently the teachers used each of the score levels on the 5-point
holistic scale. The teachers used all of the points on the scale, and the mean of the
distribution is 3.2 with a standard deviation of 1.1, which is about 25% of the range
of the scale. The reliability of scoring is .69 for a single scorer, .83 for the average
of two scorers, and .87 for three scorers, based on a generalizability analysis.

Analysis of project scores. An ANOVA was carried out for the Overall
Quality score, as well as for each of the scores for eight criteria used in judging
The Distribution of Mass Project Scores

FIGURE 10 The distribution of scores for the Mass Project.

The means of the overall Mass Project scores are shown in Figure 11 for each combination of CTBS level (low and high) and treatment (reflective-assessment and control). The most striking feature of these results is that the reflective-assessment students do significantly better than the control students on their Mass Projects, and this effect of reflective assessment appears to be greatest for students with low CTBS scores. For instance, there is little difference in mean performance between high- and low-CTBS students in the reflective-assessment classes (0.39,

---

7"Design" was not one of the criteria presented to students (see Figure 8). The teachers introduced it as a replacement for "Being Systematic" and "Being Inventive" because they felt that these were both aspects of Design and that, conversely, Design could incorporate both systematicity and inventiveness in a way that would be easier to score.

8Because projects were carried out in groups generally made up of two students, we carried out additional statistical tests in which we reduced the degrees of freedom of the error by a factor of one half to be conservative in our analyses. This would be the degrees of freedom obtained if the scores for pairs of students who worked together were always identical and the ANOVA had been carried out on only one score from each pair. This is conservative because only 33% of the students turned in duplicate reports for the Common Project and only 47% of the students turned in duplicate reports for the Chosen Project. In all cases, these reanalyses increased the probability values associated with the test by a very small amount, at most .01.
Mass Project Scores

![Graph showing Mass Project Scores]

FIGURE 11 The mean overall scores on their Mass Projects for students in the reflective-assessment and control classes, plotted as a function of their achievement level.

an effect size of 0.4 $\sigma$). In contrast, there is a large difference between their performance in the control classes (1.43, an effect size of 1.44 $\sigma$). The effect of reflective assessment, however, is independent of gender and grade (i.e., there were no interactions of reflective assessment with either of these factors).

In addition to these effects of reflective assessment, there is also a significant main effect of grade, $F(1, 106) = 4.79, p = .03$, on students' project scores. The mean score for students in Grade 7 is 2.7, whereas the mean score for Grades 8/9 is 3.2. There is no significant effect of gender on overall project scores, $F(1, 106) = 0.35$.

Mean scores for each of the separate criteria used in judging research projects are shown in Figure 12 for both reflective-assessment and control classes. The main effects of reflective assessment are significant for Design, $t(109) = 2.29, p = .01$; Reasoning, $t(106) = 2.21, p = .02$; and Teamwork, $t(69) = 2.42, p = .01$. In addition, for each of the criteria, the effect of reflective assessment is greater for the low-CTBS students than for the high-CTBS students. These results are shown in Table 1. This table reveals that, for low-CTBS students, the effect sizes range from

---

Throughout these analyses, effect sizes are calculated using the square root of the error variance. When an assessment was given as both a pretest and posttest, as was the case for the Inquiry Test, the error variance is that of the pretest.
0.25 \( \sigma \) to 1.03 \( \sigma \), whereas those for high-CTBS students range from -0.13 \( \sigma \) to 0.34 \( \sigma \). For high-achieving students, the only appreciable effects are for Teamwork (0.34 \( \sigma \) and Design (0.22 \( \sigma \). For low-achieving students, the effects are more widespread, and there are significant effects of reflective assessment (in order of magnitude) for Teamwork (1.03 \( \sigma \), Reasoning (0.79 \( \sigma \), and Design (0.77 \( \sigma \). There are also appreciable effects for Understanding Science (0.59 \( \sigma \), Using Tools (0.52 \( \sigma \), Communication (0.47 \( \sigma \), Understanding Inquiry (0.37 \( \sigma \), and Making Connections (0.25 \( \sigma \). In summary, the Reflective-Assessment Process appears to improve social interaction (Teamwork and Communication), while it also improves students' processes of inquiry (Design and Reasoning) and their understanding of science resulting from their project research.

In general, for the overall quality scores as well as the scores for each of the criteria, the effects of reflective assessment are the same: It has a positive effect on the quality of students' project work, and the effect is greater for students who enter the ThinkerTools Inquiry class with lower standardized achievement test scores.

![Mass Project Criterion Scores](image)

**FIGURE 12** The mean score on each of the assessment criteria for students' Mass Projects in the reflective-assessment and control classes.
# TABLE 1
Mean Mass Project Scores for Students at High- and Low-CTBS Achievement Levels in Both Reflective-Assessment and Control Classes

<table>
<thead>
<tr>
<th></th>
<th>Understanding Science</th>
<th>Understanding Inquiry</th>
<th>Making Connections</th>
<th>Design</th>
<th>Using Tools</th>
<th>Reasoning</th>
<th>Communication</th>
<th>Teamwork</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>σ</td>
<td>M</td>
<td>σ</td>
<td>M</td>
<td>σ</td>
<td>M</td>
<td>σ</td>
</tr>
<tr>
<td>Reflective-assessment classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High CTBS</td>
<td>3.45</td>
<td>3.57</td>
<td>3.24</td>
<td>3.54</td>
<td>3.46</td>
<td>3.52</td>
<td>3.56</td>
<td>3.94</td>
</tr>
<tr>
<td>Low CTBS</td>
<td>2.82</td>
<td>2.68</td>
<td>2.12</td>
<td>3.02</td>
<td>3.23</td>
<td>2.90</td>
<td>2.75</td>
<td>3.32</td>
</tr>
<tr>
<td>Control classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High CTBS</td>
<td>3.57</td>
<td>3.47</td>
<td>3.24</td>
<td>3.35</td>
<td>3.53</td>
<td>3.37</td>
<td>3.44</td>
<td>3.64</td>
</tr>
<tr>
<td>Low CTBS</td>
<td>2.28</td>
<td>2.32</td>
<td>1.86</td>
<td>2.38</td>
<td>2.76</td>
<td>2.17</td>
<td>2.30</td>
<td>2.42</td>
</tr>
<tr>
<td>Difference (assess-control)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High CTBS</td>
<td>-0.12</td>
<td>-0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>0.19</td>
<td>0.22</td>
<td>-0.07</td>
<td>-0.08</td>
</tr>
<tr>
<td>Low CTBS</td>
<td>0.54*</td>
<td>0.59</td>
<td>0.36</td>
<td>0.37</td>
<td>0.26</td>
<td>0.25</td>
<td>0.64**</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Note. CTBS = Comprehensive Test of Basic Skills.
*p ≤ .10. **p ≤ .05.
**Frequency of completing projects.** There are no differences between the reflective-assessment and control classes in the rates of submitting project reports (76% for the reflective-assessment classes and 73% for the control classes). The high-CTBS students in both groups had a higher rate of submitting reports (91%) than the low-CTBS students (64%). The mean achievement level of the low-CTBS students who submitted research projects is not significantly different from that of the low-CTBS students who did not. This indicates that the low-CTBS students who submitted projects are representative of that achievement level. The same is true for the high-CTBS students.

**Submitting duplicate project reports.** In their project work, students generally worked in groups, and there were cases where both students within a group submitted identical written reports, even though they were instructed to prepare individual write-ups of their joint research. (In such cases, both students were counted as having submitted duplicate projects.) As is shown in Figure 13, there was a tendency for low-CTBS students to submit duplicate projects more often than high-CTBS students in control classes, $\chi^2(1, N = 70) = 7.51, p = .006$, but this tendency was nearly eliminated when students were in reflective-assessment classes, $\chi^2(1, N = 74) = 0.53, p = .47$. In control classes, the rate of duplicate project submissions for low-CTBS students (67%) is high enough to raise questions about

![Percentage of Duplicate Reports](image)

**FIGURE 13** The percentage of duplicate reports for students in the reflective-assessment and control classes, plotted as a function of their achievement level.
their ability to independently demonstrate their understanding of the ideas and outcomes of their project. However, in classes where reflective assessment was practiced, low-CTBS students were as likely to independently show their mastery of the concepts and experimental outcomes of their collaborative project work as were high-CTBS students.

**Effects of group composition.** In forming research groups, students in the ThinkerTools classes were free to choose their partner(s). As a result, there was variability in the composition of students’ research groups which could influence students’ performance on their projects. For example, some research groups were composed entirely of high-achieving students, others were composed of low-achieving students, and still others of mixtures of high- and low-achieving students. Research by Carter and Jones (1994) indicates that this is an important factor influencing students’ performance. We therefore examined the effects of such variations in group composition on students’ scores on their research projects. Our hypothesis is that students with poorer prior educational achievement will profit from working with students with higher educational achievement, and that this advantage will be particularly strong for students in the reflective-assessment classes because reflective assessment will be modeled for the low-achieving students by the high-achieving students while they are working together.

To evaluate this hypothesis, we first classified students as low or high in their prior educational attainment on the basis of their CTBS scores (≤ 60, > 60). **Heterogeneous groups** were those containing both low- and high-CTBS students, whereas **homogeneous groups** contained only low- or only high-CTBS students. (Students who worked individually were excluded from this analysis.) A total of 106 students’ project scores were available for analysis. An ANOVA was then carried out on the Mass Project overall scores: The factors were group composition (heterogeneous vs. homogeneous), treatment (reflective assessment vs. control), and the student’s achievement level (low vs. high). Finally, planned comparisons were carried out testing the effects of heterogeneous versus homogeneous grouping on the performance of low- and high-CTBS students in the reflective-assessment and control classes. We expected that the effects of group composition would be greatest for low-CTBS students, who would benefit from working in groups with high-CTBS students. We also expected that these effects of group composition would be greatest for students in reflective-assessment classes, in which students had learned a framework for evaluating and reflecting on their work.

Our analysis confirms that there are significant benefits for students, particularly for low-CTBS students, when reflective-assessment practices are combined with heterogeneous research groups in inquiry-oriented science classes. In general, students in heterogeneous groups outperformed those in homogeneous groups when they were members of reflective-assessment classes (the mean
difference is 0.55), but not when they were in control classes (the mean difference is -0.53). In other words, there is a significant interaction of group composition and treatment, \( t(98) = 1.91, p = .03 \). However, when we looked at these effects separately for low- and high-CTBS students, we found that the benefits of heterogeneous grouping depended on students' prior level of academic achievement. In reflective-assessment classes, the mean performance of low-CTBS students was higher when they worked in heterogeneous groups than when they worked in homogeneous groups; the mean difference is 1.28 or 1.24 \( \sigma \), \( t(98) = 1.53, p = .06 \), whereas, for high-CTBS students, there was no effect of the group composition (the mean difference is -0.18). In the control classes, there were no effects of group composition on either low- or high-CTBS students' performance. Thus, in classes where reflective assessment is introduced, low-CTBS students benefit most when they participate in reflective analysis and monitoring of their work in collaboration with a higher achieving partner.

**Module 7: Chosen Research Projects**

In their Final Projects (Module 7), also known as the Chosen Research Projects, students had to choose a topic from a list of suggested topics. The projects varied in how closely they were related to the experiments students had carried out in the classroom: Some projects were clear extensions of projects included within the curriculum (e.g., investigating air resistance, mass and gravity, or projectile motion), whereas others involved exploring substantially new phenomena (e.g., investigating collisions or circular motion), and still others involved alternatives to experimental research (e.g., investigating historical figures or designing educational games). In addition, in carrying out final projects, much of their work was carried out on students' own time and over an extended period (typically several weeks). The completion of their projects thus required greater initiative on the part of students than was the case for the Module 3 projects, which were carried out as part of their in-class work. Students also worked primarily in groups, and the effects of group composition were again expected to play a roll. Each of these factors must be separately evaluated in determining the effects of reflective assessment on students' project work.

**Analysis of project scores.** Our theory is that the assessment criteria presented within the curriculum act as metacognitive tools for students to use as they reflect on the functions and results of their inquiry processes. Because students in the reflective-assessment classes have been engaged in learning and using the assessment criteria throughout the ThinkerTools curriculum, it was our expectation that their degree of success in learning reflective self-assessment should be strongly
related to the quality of their final projects. We therefore sought in our analysis of students' final projects to differentiate those students who had developed an understanding of the set of assessment concepts from those who had not. Our hypothesis is that students in the reflective-assessment classes who have learned to make distinctions among the concepts and apply them carefully in judging their work will, as a consequence, produce higher quality projects than those of students who have not learned to make such distinctions. Thus, if the reflective-assessment concepts are acting as metacognitive tools to help students as they ponder the functions and outcomes of their inquiry processes, then the students' performance should depend on how well they understand and use the assessment concepts.

We employed two measures to assess students' understanding and use of the reflective-assessment concepts: (a) the appropriateness of the evidence they cited in justifying their self-assessment scores and (b) the extent to which their pattern of scores agreed with those given by the teachers who evaluated their project. For our analysis using the first measure, the students' justifications for their self-assessment scores were rated for each of the criteria as to the relevance of the evidence they cited. These ratings were made using a 5-point scale ranging from 1 (not relevant) to 5 (clearly relevant) by a research assistant who was blind as to the student's identity. For each student, we averaged the ratings across all of the assessment criteria to obtain a mean relevance rating. We then classified students into two categories: (a) relevant evidence (Ms > 3.5) and (b) nonrelevant evidence (Ms ≤ 3.5). Finally, we looked at the quality of the students' final projects, comparing students who, according to this measure, had developed an understanding of the set of assessment concepts by the end of the curriculum with those who did not.

The results of this analysis are shown in Figure 14. Overall, students who learned to justify their self-assessment scores with appropriate evidence produced higher quality final projects than students who did not, t(41) = 1.81, p = .04. Again, we found that the importance of learning to use the assessment concepts is greatest for the low-CTBS students: There is a larger effect size for low-CTBS students (1.0 σ), t(41) = 2.10, p = .02, than for high-CTBS students (0.27 σ), t(41) = 0.54, p = .30. These findings clearly implicate the use of the assessment concepts as a tool for learning to carry out scientific inquiry, particularly for the low-CTBS students.

We conducted an additional, more extensive analysis using our second measure of students' understanding and use of the reflective-assessment concepts. This measure assesses the agreement between a student's and the teachers' scoring of a project. This is achieved by calculating the mean profile agreement (MPA) between the student's self-assessment score for each of the 10 criteria and the final criterion score agreed on by the teachers for each of the same criteria.\(^\text{10}\) The MPA for a student is given by

\(^{10}\) The criteria include those given in Figure 8 plus an evaluation of the quality of the student's own self-assessment (see the end of Appendix B). Because the teachers gave ratings for Design in place of separate evaluations for Systematicity and Inventiveness, in calculating MPA, we compared the students' evaluations of Systematicity and Inventiveness with the Design score assigned by the teachers.
MPA = \frac{1}{n} \sum_{i=1}^{n} |(s_i - \bar{s}) - (t_i - \bar{t})|,

where \( n \) is the number of criteria (10), \( s_i \) is the student's score on the \( i \)th assessment criterion, \( \bar{s} \) is the mean of the student's self-assessment scores over the 10 criteria, \( t_i \) is the teachers' moderated score for the \( i \)th criterion, and \( \bar{t} \) is the mean of the teachers' scores over the 10 criteria. MPA is the mean of the absolute differences between the student's and teachers' scores, after removing differences in the overall mean scores of the student and the teachers. It is an index of the student's ability to see and make the same distinctions among the criteria as those made by the teachers in reflecting on and judging the qualities of his or her project work. Low values of MPA indicate that the student and teacher were in close agreement in their relative scores on the different criteria, whereas high values indicate poor agreement.

For purposes of analysis, we classified students into two groups: a profile agreement group (MPA ≤ 0.6, the median score) and a profile disagreement group (with MPA > 0.6). We then looked at the effects of profile agreement on final project scores using a three-way ANOVA along with a set of planned comparisons for (a) reflective-assessment versus control classes and (b) low-achieving versus high-

![Final Project Scores](image-url)

FIGURE 14 The mean scores on the Chosen (Final) Projects for students who did and did not provide relevant evidence when justifying their self-assessment scores, plotted as a function of their achievement level.
achieving students. Factors in the analysis included: treatment (reflective assessment, control), CTBS achievement level (low, high), and profile agreement (agree, disagree). Data for a total of 60 students were available for analysis. The results are presented in Table 2.

The most important finding is that students in the reflective-assessment classes who succeeded in developing an understanding of the assessment concepts had significantly higher final project scores ($M = 3.97$) than students who did not develop such a clear understanding ($M = 2.57$). The difference in means is 1.40, which represents an effect size of 1.7 σ. Students' ability to carry out reflective assessment thus appears to have a strong influence on their ability to produce high-quality research projects. Moreover, this effect of understanding the assessment concepts

---

**TABLE 2**

Mean Scores on Final Projects for Students Who Differ in Their Understanding and Use of the Assessment Concepts

<table>
<thead>
<tr>
<th>Comparison</th>
<th>$M$</th>
<th>SE</th>
<th>$t(53)$</th>
<th>$p^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reflective-assessment classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile agreement</td>
<td>3.97</td>
<td>0.20</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Profile disagreement</td>
<td>2.57</td>
<td>0.21</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>1.40</td>
<td>3.88</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>Control classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile agreement</td>
<td>3.58</td>
<td>0.27</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Profile disagreement</td>
<td>3.09</td>
<td>0.22</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>0.49</td>
<td>1.42</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>Low-achieving students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile agreement</td>
<td>3.71</td>
<td>0.30</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Profile disagreement</td>
<td>2.24</td>
<td>0.22</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>1.47</td>
<td>3.27</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>High-achieving students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile agreement</td>
<td>3.84</td>
<td>0.18</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Profile disagreement</td>
<td>3.41</td>
<td>0.21</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>0.43</td>
<td>1.52</td>
<td>.07</td>
<td></td>
</tr>
</tbody>
</table>

$a$One-tailed test.

---

Sixty students might seem like a low sample size for a study that started with 343 students. The following explains how this occurred. As discussed previously, projects from 8 of the 12 classes were considered (matched treatment and control classes for each teacher at each grade level)—reducing the sample size to 221. Twenty-seven percent of these students did not turn in Chosen Research Projects—reducing the sample size to 161. Due to limitations in the time and resources available for the scoring study (described previously), not all of these projects were scored—reducing the sample size to 116. CTBS scores were missing for many of the students (due to their being absent on the day of the CTBS test or to lack of parental permission for us to obtain the score)—reducing the sample size to 93 students. Finally, the students' self-assessments were missing for 33 of these projects (students often did not turn them in with their projects)—reducing the sample size to 60.
seems to depend on the criteria having been learned and applied over the full course of the ThinkerTools curriculum. Students in the control classes were also asked to assess their final projects using the scoring criteria. For these students, this was the first time they had seen the scoring criteria and the first time that they had been asked to do a reflective self-assessment. The effect of understanding the criteria on the quality of their final projects is much smaller for these control students (0.49 or 0.6 σ) and is only marginally significant. This means that the effect of understanding the criteria cannot be attributed simply to students' ability to interpret their work when they are given a set of criteria. Rather, it appears to be due to students learning the reflective-assessment criteria and applying them over an extended time as they carry out science inquiry projects throughout the curriculum.

Nonetheless, what about the possibility that students who learn how to use the assessment criteria may have generally higher academic ability than those who do not, and that this is responsible for their producing higher quality projects? First, use of the factorial ANOVA means that, in looking at the relationship between any two factors (e.g., treatment and profile agreement), we are controlling for the third factor (e.g., achievement level) through equal weighting of means for each level of that factor. Second, there is no significant association between CTBS achievement level and profile agreement, \( \chi^2(1, N = 60) = 2.40, p = .12, \phi = -0.23 \). So, the results are not an artifact of students who successfully learn to use the criteria having generally higher achievement.

What then are the particular effects of learning the reflective-assessment criteria for low- and high-achieving students? Our theory is that high-achieving students already have implicit metacognitive skills for reflection, whereas low-achieving students lack such implicit skills (see also Campione, 1987; Carey, 1985; Feuerstein et al., 1985; Nickerson et al., 1985). The benefits of learning reflective assessment should therefore be greater for low-achieving students than for high-achieving students.

To address this possibility, we examined separately the effects of profile agreement for low- and high-achieving subgroups of students (see the bottom of Table 2 and Figure 15). It can be seen that, for low-achieving students, developing an understanding of the criteria and their use has a large impact on their project scores. When such students show high profile agreement, their mean score on their final projects is 3.71, compared with a mean of only 2.24 for those who show low profile agreement (the mean difference is 1.47 or 1.8 σ, and is significant). In addition, their performance is close to that of high-achieving students who have also learned to use the criteria (3.84). For high-achieving students, the effects of profile agreement are smaller, with a mean of 3.84 for those who have understood how to use the criteria and 3.41 for those who have not (the mean difference is 0.43 or 0.5 σ, and is only marginally significant). These results are consistent with our hypothesis that high-achieving students already have metacognitive skills that help them learn how to carry out inquiry projects, even
if they have not focused on learning the particular assessment criteria we have introduced in our curriculum. In contrast, for lower achieving students, including curricular activities to help them learn how to think critically about their work is more crucial because, in the absence of an explicit teaching of these skills, such students have less well developed metacognitive skills to rely on when learning how to carry out science research projects. However, it is clear that learning to use the assessment criteria in reflecting on work produces gains in the quality of research projects for both high- and low-achieving students.

**Frequencies of completing projects.** A measure of students’ interest and confidence in their work is their rate of completing their projects and turning in written research reports. We therefore tallied the percentage of students who turned in final projects for both the reflective-assessment and control classes. We found that there is a significant difference of 12% in rates of submitting completed projects between the reflective-assessment and control students, $\chi^2(1, N = 221) = 4.6, p = .03$, with 79% of the reflective-assessment students turning in final projects and 67% of the control students doing so. This effect of reflective assessment on turning in projects occurred for all grade levels and CTBS achievement levels, although

![Final Project Scores](image-url)
the effect was larger for Grade 7 (20%) than for Grades 8/9 (6%)\textsuperscript{12} and was larger for the low-CTBS students (16%) than high-CTBS students (6%).\textsuperscript{13} So, it is a very general effect. These results support the value of the Reflective-Assessment Process, particularly for the younger grades and for the low-CTBS students.

Comparing the submission rates for their Chosen Project with those for the Common Project reveals that the rates remain the same for the high-CTBS students (averaging about 90% in both cases). There is, however, a difference for low-CTBS students: Their average submission rate for the Common Projects was 64%. For the Chosen Project, their submission rate increased in the reflective-assessment classes (to 71%), whereas it decreased in the control classes (to 55%). Because completing Chosen Projects requires greater initiative and interest on the part of students, this comparison offers further evidence of the value of reflective assessment for the lower achieving students.

\textit{Submitting duplicate project reports.} Students in the reflective-assessment classes had a lower rate of submitting duplicate Chosen Project Reports (43%) than those in the control classes (51%), although this difference is not significant. The teachers did not insist that students write these Chosen Project Reports independently of their research partners, and this probably accounts for the higher percentage of duplicate reports than occurred for the Common (Mass) Project. It should be noted that when professional researchers collaborate, they typically produce joint reports (like this one) rather than individual write-ups. In this instructional context, however, individual write-ups are more useful for assessment and research purposes.

\textit{Choices of project topics.} We found that, for the final projects, there were differences between reflective-assessment classes and control classes in the nature of the topics students chose to undertake. The project topics were classified into three categories: (a) familiar topics that represent extensions of projects that have already been encountered in the curriculum, (b) novel projects that involve physical phenomena that have not already been explored in the curriculum, and (c) miscellaneous projects that do not involve designing and running experiments (historical projects and educational game projects). We tallied the frequencies with which low- and high-CTBS students in each of the treatment groups selected projects of each

\textsuperscript{12}Students in Grade 7 turned in work at rates of 81% in the reflective-assessment classes and 61% in the control classes, compared with 78% and 72%, respectively, for students in Grades 8/9.

\textsuperscript{13}High-CTBS students generally turned in more completed work, at rates of 92% in the reflective-assessment classes and 86% in the control classes, compared with 71% and 55%, respectively, for the low-CTBS students.
type. The effect of the Reflective-Assessment Process was to increase the percentage of low-CTBS students attempting to take on the more difficult novel projects, from 20% (in the control classes) to 34% (in the reflective-assessment classes), although this difference is not significant.

**Effects of group composition.** In an ANOVA of overall Final (Chosen) Project scores, we again found that it is beneficial for students to work in heterogeneous research groups. For students in classes where they used reflective assessment, the composition of the research groups was extremely important in determining how well the low-CTBS students (but not the high-CTBS students) performed, as it was in our analysis of Mass Project scores. For low-CTBS students in the reflective-assessment classes, we found a significant effect of group composition on their project scores, with scores for heterogeneous groups exceeding those for homogeneous groups by an average of 0.85 (0.79 σ), $t(71) = 1.66, p = .05$. In contrast, we did not find any effects of group composition for low-CTBS students in the control classes (the mean difference is 0.18), which suggests that the effects of grouping are due to reflective-assessment conversations students have about their research as they carry out their projects. Again, there were no significant effects of group composition for high-CTBS students in either reflective-assessment classes (with a mean difference of 0.13 between those in heterogeneous vs. homogeneous groups) or control classes (with a mean difference of 0.11). Presumably, high-achieving students are able to reflect on their performance regardless of the composition of their research groups and are not impeded by working with lower achieving students. Together, these results corroborate our earlier interpretation that low-CTBS students benefit from participating in reflective analysis and monitoring of their work in collaboration with their higher achieving partners.

**Inquiry Test**

Up until this point, we have been presenting findings from the collaborative research projects carried out within the curriculum. Here, we present the findings from another measure of students' inquiry expertise—mainly, the Inquiry Test. This is a written test, administered in a 45-min period, in which the students work individually. In this test, the students are given a research question (i.e., “What is the effect of varying the weight of an object on what sliding friction does to its motion?”) and are asked to develop some alternative hypotheses for what they think might happen. They then have to design an experiment that would let them determine what actually happens. Next, they have to make up some measurements that they think they might get if they were to actually carry out their experiment. Finally, they have to analyze their

---

14 A total of 79 scored students' projects were available for the analysis of group composition.
made-up data to reach a conclusion about the answer to the research question and relate their conclusion back to their original hypotheses.

The Inquiry Test was given as a pretest and posttest in one of our schools, where it was taken by students in five ThinkerTools classes, three in Grade 7 and two in Grade 9. Unfortunately, pretest and posttest scores were not available for the second school, which began the curriculum before the inquiry test had been developed. We analyzed the data for students for whom we had both pretest and posttest scores and for whom we also had CTBS scores (n = 90).

**Scoring.** The Inquiry Test was scored solely with regard to inquiry skills. In other words, students were assessed on their understanding of the inquiry process and not for their conceptual understanding of physics. Thus, whether or not the students' hypotheses and conclusions embodied the correct physics was treated as irrelevant. The test was scored analytically by coding discrete features of the students' responses with regard to inquiry (such as whether or not they varied mass in their experiment). In this coding, the scorer was blind as to the students' gender, grade, whether the test was a pretest or posttest, and whether the student was in the reflective-assessment or control group. For each of the four sections of the test, separate scores were assigned. The scoring categories (along with the weights used in determining a combined score) are Hypotheses (15%), Experiment (35%), Results (15%), and Conclusions (15%). In addition, a fifth score representing Coherence (20%) was determined by judging the consistency of the students' experiment with their hypotheses, their results with their experiment, and their conclusions with their results, experiment, and hypotheses. The Overall Inquiry score was defined as the weighted combination of these five subscores. The range of possible Overall Scores was 0 to 100. The internal consistency reliability of the Inquiry Test is 0.88 for the pretests (coefficient α for the five subscores) and 0.86 for the posttests. The interscorer reliability for the five subscores (based on having 10% of the tests scored by two research assistants) ranges from 0.81 (for Hypothesis) to 0.94 (for Conclusions). The interrater reliability for the weighted average of these subscores is 0.96.

**Results.** We first carried out an ANOVA, which included time of testing as a repeated measures factor. As expected, the overall scores on the inquiry test improved significantly, $F(1, 79) = 30.7, p < .001$. The average gain was 15 points, which, measured in terms of the standard deviation of the pretest score distribution (19), represents an effect size of 0.8 $σ$. Although there was a large effect of instruction, the average gain scores did not differ significantly as a function of either grade or gender. So, the curriculum was equally effective for students at the different grade levels as well as for male and female students.

In addition to this main effect of instruction, we found that the reflective-assessment treatment also had a significant effect on students' gain scores, $F(1, 79) = 6.9,$
Overall, students who practiced reflective assessment had an average gain score of 23 (improving from 40 to 63) or 1.2 σ, whereas students who did not practice reflective assessment had an average gain of only 8 (improving from 46 to 54) or 0.4 σ. Thus, the average gain for students in the reflective-assessment classes is three times that for control classes.

An even more interesting result is that the low-CTBS students show greater improvements in inquiry scores when they engage in reflective assessment than do the high-CTBS students. Figure 16 shows the average pretest and posttest scores on the Inquiry Test for students in the reflective-assessment classes compared with those in the control classes, for both high- and low-CTBS students. For low-CTBS students, the average gain from pretest to posttest was 25.6 (1.3 σ) in reflective-assessment classes and 6.2 (0.3 σ) for control classes; the difference between these average gains is significant, t(79) = 2.25, p = .02. In contrast, for the high-CTBS students, the average gain scores were 18.7 (1.0 σ) in the reflective-assessment classes and 9.9 (0.5 σ) in the control classes; this difference is marginally significant, t(79) = 1.45, p = .08. Thus, the Reflective-Assessment Process, although beneficial for all students, has the effect of bringing the low-CTBS students up to a level that is closer to that of the high-CTBS students. The average posttest scores for the high- and low-CTBS students were 72 and 53, respectively, for the reflective-assessment classes (a difference of 19 or 1.0 σ), whereas they were 68 and 39, respectively, for the control classes (a difference of 29 or 1.5 σ). Because

**Figure 16** The mean scores on the Inquiry Test for high- and low-achieving students in the reflective-assessment and control classes.
Inquiry Gain Subscores

FIGURE 17 Average gains on the Inquiry Test subscores for students in the reflective-assessment and control classes.

the highest score that anyone got on this test was 89 (out of 100), these results are unlikely to be due to a ceiling effect restricting the gains for the high-CTBS students.

For each of the subscores on this test, the average gains for students in the reflective-assessment classes exceed those for the control classes, as shown in Figure 17. It is interesting that the largest effects of reflective assessment are on the most difficult aspects of the test, namely those related to the Results, Conclusions, and Coherence subscores. These require students to make up data that are appropriate to their experimental design, to analyze their data in order to reach a conclusion that is consistent with their data (students often reach conclusions based on their favorite hypothesis and ignore the data; Chinn & Brewer, 1993; D. Kuhn, 1988), and to maintain coherence. The Coherence subscore measures the extent to which the students’ experiments follow from their hypotheses, their data relate to their experiment, their conclusions follow from their data, and so on. Creating a research study with this kind of coherence is an important measure of inquiry expertise, and it is on this measure that students showed the largest differential gain for the reflective-assessment classes, amounting to 1.0 σ, compared to 0.2 σ for the control classes.

15Planned contrasts (one-tailed) were carried out comparing the differences between mean gain scores for the reflective-assessment versus control classes. The contrasts showed that there is no significant difference for the Hypothesis score, t(80) = 0.66, p = .25. There is a marginally significant difference for the Experimental Design score, t(80) = 1.59, p = .06. There are significant differences for the Results score, t(80) = 2.54, p = .007; Conclusions score, t(80) = 1.88, p = .03; and Coherence score, t(80) = 2.82, p = .003.
Effects of understanding the assessment criteria. As in our analysis of Final (Chosen) Project scores, we carried out a second ANOVA to examine the effects of students’ learning to use the assessment criteria on the growth of their inquiry skills. Our measure of inquiry learning is the gain in overall scores on the Inquiry Test. As in our earlier analysis, we use the MPA score as our measure of a student’s understanding of the inquiry criteria. Our main finding is that the effect of participating in reflective-assessment classes (as compared with control classes) is greatest for students who show a knowledge of the assessment criteria and can use them to distinguish among aspects of their performance addressed by the criteria. For students who show high profile agreement with the teachers, the mean Inquiry Test gain scores are 29.7 for the reflective-assessment classes and −1.1 for the control classes, a difference of 30.8, t(35) = 3.30, p = .001. For students who show poor profile agreement with the teachers, the mean gain scores are 18.5 for the reflective-assessment classes and 7.3 for the control classes, a difference of 11.2, t(35) = 1.31, p = .10. Thus, we again see that the reflective-assessment effects are greatest for students who have learned to use the assessment criteria to reflect accurately on the qualities of their work.

Discussion of the Reflective-Assessment Effect

The results obtained for the various inquiry measures—the measures of performance on research reports and the gain scores on the Inquiry Test—show a consistent pattern: The Reflective-Assessment Process is having a strong positive effect on students’ inquiry performance, and this effect is greatest for students who have low scores on a standardized achievement test (CTBS) and for students who have developed a clear understanding of the assessment criteria and how to use them in reflecting on their work.

Why is reflective assessment effective? Our conjecture is that the scoring criteria (which are introduced to students as “guidelines for judging your work”) help students to learn about the desired process and products of research. This learning gives students greater confidence that they know what is required and that they can accomplish it. Also, knowing that their research is going to be assessed not only by themselves but also by their teacher and fellow classmates may motivate them to try harder (students in control classes knew their research would be graded only by the teacher). Finally, the process of reflecting on how well they did their research and then having a chance to try again in the next module provides feedback and opportunity for continual improvement in their research. This “reflect-and-try-again” cycle is repeated seven times in the curriculum. Such an opportunity for continual reflection and improvement, which our findings indicate are beneficial to all students and particularly for low-achieving students, is missing in most
curricula (Collins, 1994). These results are in agreement with the general design recommendations of Frederiksen and Collins (1989) for building a systemically valid assessment system, that is, a system which acts to improve the cognitive skills it is measuring.

**Why is the reflective-assessment effect greater for low-CTBS students?**

Our research indicates that the Reflective-Assessment Process is effective for all students, but particularly for low-achieving students. Our hypothesis is that a major factor causing students to be low achievers is that they lack the metacognitive skills of monitoring and reflecting on their work. Thus, if they learn these metacognitive skills, they can perform as well as high achievers. Our results offer support for this hypothesis: Low-CTBS students in reflective-assessment classes did about as well as high-CTBS students on their research projects.

Some important clues as to how this learning occurs lie in the following findings:

1. **Learning to use the assessment criteria matters:** Students in reflective-assessment classes—particularly low-CTBS students—who learn to use the assessment criteria to make criterion-related distinctions about their work have higher scores on their projects and on the Inquiry Test than those who do not learn to make such distinctions using the criteria.

2. **Reflective assessment increases motivation:** The rate of turning in Final Project Reports is higher for students in the reflective-assessment classes than in the control classes, particularly for the low-CTBS students; also, the low-CTBS students in the reflective-assessment classes are more frequently choosing the difficult research topics.

3. **Reflective assessment increases confidence in one's ability to perform on one's own:** The incidence of turning in reports that are identical to their research partner's is less for students in the reflective-assessment classes than in the control classes, particularly for the low-CTBS students.

4. **Reflective assessment fosters communication and collaboration, which improve learning:** The scores of low-CTBS students in the control classes were independent of whether they collaborated with low- or high-CTBS students, whereas in the reflective-assessment classes, the scores of the low-CTBS students were higher if they collaborated with high-CTBS students. In addition, the Teamwork scores were higher for students in the reflective-assessment classes (as shown in Figure 12) than in the control classes.

Thus, it is the combination of collaborating with a high-achieving partner and engaging in the Reflective-Assessment Process that appears to be effective for the low-achieving students. The collaborative nature of the research groups appears to be a key factor in making self-assessment a valuable activity, and the Reflective-Assessment Process appears to foster that collaboration (cf. Carter & Jones, 1994;
Slavin, 1987; Vygotsky, 1978). It gives students an opportunity to reflect on their work as well as a vocabulary to use in these discussions. Our conjecture is that high-CTBS students have more highly developed metacognitive monitoring and reflecting skills at the start of the curriculum (which is the primary reason they are high achievers). As they work to incorporate the Inquiry Cycle and the Reflective-Assessment Criteria into their prior knowledge of metacognition and inquiry, they provide a model of how to monitor, reflect on, and revise one's inquiry skills for their low-CTBS partners as they collaborate to design and carry out their research projects. As a brief example, consider the following interchange between a high- and low-CTBS student who are collaborators. They are working on the self-assessment page in their research books shown in Figure 9.

Low-achieving student: What did you put for "being systematic?"
High-achieving student: Well, at the beginning we were very systematic and did everything really carefully, but as it got closer to the end and we had less time it started to get loose and sloppy, so I give us a 3.

Low-achieving student: I think we communicated pretty well don't you?
High-achieving student: I gave myself a 4 on that because I think we did a good job of discussing everything together, and I think my write up of our experiment is clear.

Low-achieving student: Yeah, I think I'll give myself a 3 because my writing isn't good like yours.

The videotapes indicate wide variation in how such self-assessments are carried out. They vary from each student working on them independently, to one student copying what the other has written, to including conversations like the one just presented. A more extended process analysis of pairs of high- and low-CTBS students working together would help unravel what is happening in these collaborations and how it is benefiting both students, particularly the low-CTBS students.

Such reflective-assessment activities carried out at the end of each step in the Inquiry Cycle and at the end of each module provide many opportunities for the low-CTBS students to practice self-assessment. This, in conjunction with the scaffolded inquiry activities, enables the low-CTBS students to develop confidence and skills to do better research, as well as to develop important metacognitive skills that will help them in subjects other than science.

**RESULTS: PHYSICS KNOWLEDGE**

We turn now to the conceptual understanding of physics that the students have developed. Here, we address two main research questions:
1. How well did students develop an understanding of the physics of force and motion?
2. How did their knowledge of physics depend on their mastery of the process of inquiry?

In answering the first question, we are interested in the extent to which an inquiry-based science class in which students are treated as researchers can be relied on to develop appropriate knowledge of the relevant science. While exploring this question, we shall also consider the effects of reflective assessment, achievement level, gender, and grade level on the learning of physics. In addressing the second question, we are attempting to validate our hypothesis that students' inquiry activities are instrumental in enabling them to construct their models of force and motion in such classes.

We begin by considering evidence of the quality of physics knowledge acquired by students in the ThinkerTools classes. At the end of the course, the students were given two paper-and-pencil tests of their knowledge of the physics addressed in the curriculum. The first, called the Conceptual Model Test, presents problems involving the behavior of objects (called dots) in situations similar to those encountered in the computer simulations. It is designed to assess students' understanding of the conceptual model we hoped they would develop. The second, called the Applied Physics Test, presents problems concerning the behavior of objects in the real world. This test was also given as a pretest. In prior research, items on this test have been given to students in a variety of other curricular contexts, so we can compare the performance of our current students with that of those reference groups. For each test, we carried out the standard ANOVA with treatment (reflective assessment vs. control), gender, CTBS score level (low ≤ 60 and high > 60), and grade (7 vs. 8/9) as factors. In addition, in the analyses of the Applied Physics Test, we included a repeated measures factor (pretest vs. posttest) to assess the instructional effects of the ThinkerTools Inquiry Curriculum.

Conceptual Model Test

In the Conceptual Model Test, students have to answer questions about the behavior of objects (called dots) in the computer model. This test is designed to assess students' understanding of the conceptual model we hoped they would develop from their experimental activities. It contains 29 items of three general types:

- **Representation problems** that assess the students' understanding of the representations needed for the conceptual model (see Figure 18 for an example).
- **Prediction problems** that ask students to apply their conceptual model to make predictions about the effects of forces (i.e., "What would happen if ...?") questions such as that shown in Figure 19).
Suppose that the datacross went from looking like it does in datacross(a) to looking like it does in datacross(b).

Which statement best describes what happened to the speed of the dot?

(a) It decreased.
(b) It increased.
(c) It stayed the same.

Which dot prints best show what happened to the motion of the dot?

(a) •
(b) •
(c) •

FIGURE 18 A sample representation problem from the Conceptual Model Test.

• Engineering problems that ask students to show what forces would have to be applied to make the dot exhibit a particular behavior (i.e., “How would you achieve ... ?” questions such as that shown in Figure 20).

Because the test uses terms and representations introduced in the curriculum (e.g., dots, dotprints, and datacrosses), it was not given as a pretest. The student’s score on the test is the number correct out of those attempted. The Spearman–Brown estimate of the test reliability based on a split-half correlation is 0.87.

Results. On the Conceptual Model Test, the students’ average total score was 63% correct with a standard deviation of 16%. The students’ performance
Suppose that the dot is stopped.
You give it an impulse upward.
Then you give it an impulse to the right.
Then you give it an impulse downward.
There is no friction or gravity.

Circle the path the dot would make.

FIGURE 19 A sample prediction problem from the Conceptual Model Test.

Suppose that you wanted the dot to make the shape shown below.
Draw in arrows to show the direction of the impulse that you
would give at each point shown on the path.
You can only give one impulse at each point.

FIGURE 20 A sample engineering problem from the Conceptual Model Test.
was the same for all three teachers and for all three grade levels, seventh through ninth. It was also about the same as that of sixth-grade students who participated in an earlier study in which we evaluated an earlier version of our curriculum (B. White, 1993b).

The ThinkerTools students did better (on the single item we have in common from another earlier study, see B. White; 1981, 1983) than a comparison group of 40 high school students who had taken a traditional PSSC physics course, \( \chi^2(1, N = 319) = 4.52, p = .03 \). On this difficult engineering item, shown in Figure 20, 44% of the ThinkerTools students drew the correct sequence of impulses, whereas only 25% of the high school physics students drew the correct sequence. In addition, the ThinkerTools students do as well on this item as high school students who have taken the PSSC course and who have also interacted with some of the ThinkerTools computer simulation activities designed to help students understand Newtonian physics, \( \chi^2(1, N = 323) = 0.50, p = .48 \) (B. White, 1981, 1984). Again, these results are very encouraging because the high school physics students were in a more advantageous educational situation: They were an older, higher achieving group taught by an experienced physics teacher (who had studied physics in college) in a suburban public school with much smaller class sizes (averaging close to 20 instead of 30).

An ANOVA of our students’ performance on the Conceptual Model Test revealed a main effect of CTBS level, with an average of 76% correct for high-CTBS students and 49% for low-CTBS students, and a main effect of gender, with an average of 70% correct for male students and 56% for female students. In addition, there was a significant interaction of gender with grade, \( F(1, 241) = 7.98, p < .005 \). The gender difference (male student scores minus female student scores) is larger at Grades 8/9 where it is 20% (1.2 \( \sigma \)), than at Grade 7 where it is only 7% (0.5 \( \sigma \)). This finding is consistent with the gender effect seen on the Applied Physics pretest (presented in the next section), which suggests that the gender differences may represent performance differences that existed at the start of the curriculum. There is, however, an interesting exception: Students of one of the teachers, teaching at the seventh-grade level, did not show a gender effect on the Conceptual Model Test. This teacher’s teaching style when she taught the ThinkerTools curriculum was able largely to overcome the gender gap (it should be noted that the gender gap on the Applied Physics Test is lowest at the seventh grade, so there is less of a gap to overcome at this grade level). This teacher has what could be characterized as an encouraging, gentle approach to teaching the ThinkerTools Inquiry Curriculum, in which she consistently exhibits a genuine interest in the students’ ideas in a supportive manner. She rarely criticizes students’ theories or experiments (relying on the students to discover the problems for themselves), and she gives frequent praise to students for sharing their ideas with the class. These aspects of her teaching style may be responsible for her success with both male and female students.
TABLE 3
Average Scores on the Conceptual Model Test by CTBS Level, Grade, and Treatment

<table>
<thead>
<tr>
<th></th>
<th>Grade 7</th>
<th></th>
<th></th>
<th>Grades 8/9</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assess</td>
<td>Control</td>
<td>Δ</td>
<td>σ</td>
<td>Assess</td>
<td>Control</td>
</tr>
<tr>
<td>Low CTBS</td>
<td>53.1</td>
<td>44.5</td>
<td>8.6 0.5</td>
<td>74.9 79.0</td>
<td>49.5</td>
<td>49.8</td>
</tr>
<tr>
<td>High CTBS</td>
<td>74.9</td>
<td>79.0</td>
<td>-4.0 -0.2</td>
<td>76.3 75.4</td>
<td>76.3</td>
<td>75.4</td>
</tr>
</tbody>
</table>

Note. CTBS = Comprehensive Test of Basic Skills.

*Significant with p < .02.

Treatment effects. In our initial ANOVA, there is evidence for an effect of reflective assessment on the Conceptual Model Test total scores of low-CTBS students, although these are restricted to the lower grade (i.e., Grade 7). Average scores for high- and low-CTBS students in the treatment and control groups are given in Table 3 for each grade level. There is a significant improvement in the scores of reflective-assessment students over control students (8.6% or 0.5 σ) for seventh-grade low-CTBS students. The result is a reduction in the score difference between high-CTBS and low-CTBS seventh-grade students (high minus low) from 35% (2.1 σ) in the control classes to 22% (1.3 σ) in the reflective-assessment classes. However, there are no significant improvements associated with reflective assessment for students in the higher grades (Grades 8/9) or for high-CTBS students in any grade (whose average score was 75% or greater in all groups). The effects of reflective assessment are thus consistent with their effects on inquiry test scores but are limited to the younger and lower achieving students.

Effects of understanding the assessment criteria. As in our analysis of Final Project and Inquiry Test scores, we carried out a second ANOVA that included profile agreement (agree, disagree) as well as treatment (reflective assessment, control) and CTBS achievement level (low, high) as factors. Our hypothesis was that students’ mastery of the conceptual models for force and motion would be greatest (a) for students in the reflective-assessment classes and (b) for students who developed a clear understanding of the assessment criteria and how to use them in reflecting on their research projects. We used the MPA score as our index of students’ understanding and use of the assessment criteria. In this analysis, the sample was smaller (n = 60) because MPA scores were only available for a subset of the students as explained in footnote 11. The results of this analysis are shown in Table 4.

With regard to students’ success in understanding the conceptual model of force and motion, Table 4 shows that there are significant main effects of the reflective-assessment treatment, CTBS level, and a marginally significant effect of profile
<table>
<thead>
<tr>
<th>Comparison</th>
<th>M</th>
<th>SE</th>
<th>t(54)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflective assessment</td>
<td>69.8</td>
<td>3.2</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Control</td>
<td>61.4</td>
<td>3.8</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>8.4</td>
<td>1.70</td>
<td>—</td>
<td>.05</td>
</tr>
<tr>
<td>CTBS achievement level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>76.2</td>
<td>3.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Low</td>
<td>55.0</td>
<td>4.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>21.1</td>
<td>4.20</td>
<td>—</td>
<td>.000</td>
</tr>
<tr>
<td>Profile Agreement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agree</td>
<td>69.3</td>
<td>3.8</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Disagree</td>
<td>61.9</td>
<td>3.4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>7.4</td>
<td>1.46</td>
<td>—</td>
<td>.07</td>
</tr>
<tr>
<td>Treatment × CTBS interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflective assessment, high</td>
<td>73.3</td>
<td>4.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Reflective assessment, low</td>
<td>66.3</td>
<td>5.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>7.0</td>
<td>1.07</td>
<td>—</td>
<td>.13</td>
</tr>
<tr>
<td>Control, high</td>
<td>79.1</td>
<td>4.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Control, low</td>
<td>43.8</td>
<td>6.2</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Difference</td>
<td>35.3</td>
<td>4.60</td>
<td>—</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note.  CTBS = Comprehensive Test of Basic Skills.

*One-tailed test.

agreement. Thus, we have additional evidence that developing an ability to reflect on one's work has a positive impact on learning science content as well as on learning how to do inquiry. Note that, when we control for differences in students' understanding of the inquiry criteria (by including profile agreement as a factor in our analysis), we find that there is also a significant interaction of treatment with CTBS achievement level, \( F(1,54) = 7.87, p = .007 \). The pattern of results (shown in Figure 21) is similar to what we found in analyzing students' scores on research projects and on the Inquiry Test. Again, we find that learning to use the assessment criteria has had a pronounced effect on the performance of low-achieving students. For these students, the mean score on the Conceptual Model Test is only 43.8 for the control classes, whereas it is 66.3 for the reflective-assessment classes. Thus, we have evidence that when low-achieving students have the benefit of learning reflective assessment, their knowledge of the physics models developed in the curriculum begins to approach that of higher achieving students.

**Representation, engineering, and prediction problems.** We found that performance on representation problems was higher than that on engineering and
prediction problems. The average score was 71% correct for representation problems, whereas it was 62% correct for engineering problems and 63% correct for prediction problems. Thus, the curriculum and software appear to have been particularly successful in conveying to students the meanings of representations such as the datacross and dotprints and how to use them in reasoning about force and motion. Finally, although the effects of CTBS level (the average for high minus low students) were equivalent for the three types of problems (26% or 1.2 $\sigma$ for representation, 29% or 1.4 $\sigma$ for engineering, and 26% or 1.4 $\sigma$ for prediction), the effects of gender (the average for male minus female students) were greater for the engineering problems (21% or 1.0 $\sigma$) than for the representation problems (12% or 0.6 $\sigma$) and prediction problems (11% or 0.6 $\sigma$). An example of an engineering problem was given in Figure 20. Such items require students to answer “How would you make this happen?” questions as opposed to “What would happen if ...?” questions. The differential performance may be due to an attitudinal difference, such as female students being less interested in problems that relate to engineering and control as opposed to prediction and explanation. It may also be due to the two-dimensional, spatial reasoning that can be applied in their solution. Alterna-

![Conceptual-Model Test Scores](image-url)

**FIGURE 21** The mean scores on the Conceptual Model Test for students in the reflective-assessment and control classes, controlling for their understanding of the assessment criteria and plotted as a function of their achievement level.
tively, it may simply reflect the fact that “How would you make this happen?” is the problem you solve when engaging in the game-like inquiry activities, such as those shown in Figures 3 and 4, and male students may be more interested in and thus benefit more from these game-like computer activities.

Applied Physics Test

We next turn to the results for the Applied Physics Test, in which the students had to apply their knowledge of physics to solve problems in real-world situations. This test consists of items like those shown in Figures 22, 23, and 24. In the question shown in Figure 23, for example, the students are asked to predict the trajectory of a ball that is kicked off a cliff and to explain why it would take that path. Items on the Applied Physics Test were derived from prior research investigating students’

Imagine that you drop two identical balls from different heights. Both balls are dropped at exactly the same time.

![Image of two balls](image)

Which ball hits the floor first? (circle your choice below)
(A) The lower ball.
(B) The higher ball.
(C) Both balls hit the floor at the same time.

Explain the reasons for your choice: ________________________

Which ball is going faster when it hits the floor? (circle your choice below)
(A) The lower ball.
(B) The higher ball.
(C) Both balls are going the same speed when they hit the floor.

Explain the reasons for your choice: ________________________

FIGURE 22 A sample gravity problem from the Applied Physics Test.
Imagine that you kick a ball horizontally (→) off a cliff. Drawn below are three paths that someone might think the ball would take as it falls to the ground.

A B C

Circle the path you think is correct: A B C

Explain the reasons for your choice: ____________________________

______________________________

FIGURE 23 A sample trajectory problem from the Applied Physics Test.

Imagine that you are running along, and as you are running, you want to drop a ball so that it lands in a bucket below. You do this without stopping, and you just let go of the ball — you don't throw it.

At which point should you drop the ball so that it lands in the bucket?

(circle your choice): A B C

Explain the reasons for your choice:

______________________________

______________________________

FIGURE 24 A sample transfer problem from the Applied Physics Test.
understanding and misconceptions in the domain of force and motion (Clement, 1982; diSessa, 1982; McCloskey, 1983; McCloskey, Caramazza, & Green, 1980; McDermott, 1984; Minstrell, 1982; B. White, 1981, 1993b). These items have also been used extensively in our earlier studies of the role that computer simulations can play in learning physics (B. White, 1981, 1984, 1993b).

While the problems provide multiple choices for responses, the students are also asked to provide written explanations for their choices. Two scores were derived from their responses: (a) a multiple-choice score, which is the percentage of correct responses for those items attempted; and (b) an explanations score, which is based on an analysis of the correctness of the physics employed in students’ explanations. A research assistant with an undergraduate degree in physics assigned scores using a rubric having the following categories: 0, when students gave an incorrect explanation; 1, when they gave a correct but incomplete explanation in which they mention only a single variable and may have used physics concepts imprecisely; 2, when they gave a more elaborated explanation including several variables and used all physics concepts correctly; and 3, when students gave a full explanation of all of the factors and their causal interactions and used the physics concepts precisely. Exemplars for each score level were separately developed for each item so that scoring would be as tightly constrained as possible.

Reliability. The interscorer reliability for the explanations scores is .91 based on duplicate scoring of 10% of the tests (the second scorer was a research assistant with an undergraduate degree in astronomy). We also determined estimates of the test’s reliability by applying the Spearman–Brown formula to correlations of odd and even item totals. Reliabilities were calculated separately for both pretest and posttest scores. For multiple-choice scores, they were .42 for the pretest and .75 for the posttest. The low pretest reliability suggests that students are not applying consistent laws of physics to the different items (either Newtonian, Aristotelian, or other). They either have situation-specific rules, or are switching theories from question to question, or are simply guessing. For the explanations scores, the test’s reliabilities were .73 for the pretest and .84 for the posttest. For the pretests, the test’s reliability was thus lower for multiple-choice scores than for explanations. This may be partly due to the fact that multiple-choice responses could be based on pure guesses, whereas explanations can only be written when students have a reason for their choice. The reliabilities of the two posttest scores are more comparable because, in this case, students were more apt to be using the Newtonian physics they learned to make their responses.

Multiple-Choice Scores

We will consider first the results for the multiple-choice scores. The total number of tests available for analysis was 237. The main finding is that students showed...
large improvements as a result of instruction, $F(1, 226) = 270.8, p < .001$, from an average of 42% correct on the pretest to 68% on the posttest. Their average gain was 26%. Measured in terms of the standard deviation of the pretest score distribution (17%), this represents an effect size of 1.5 $\sigma$. However, unlike the results for the Conceptual Model Test, there are no significant effects of the reflective-assessment treatment. The advantages of reflective assessment thus appear for measures of scientific inquiry and for science knowledge assessed in the context of the computer model, but they do not show up on more applied physics problems.

In the ANOVA, there is also a significant interaction of CTBS level with instruction, $F(1, 226) = 35.8, p < .001$, with the size of students’ gains depending on their prior achievement level on the CTBS. Average pretest scores for the low- and high-CTBS students were 40% and 44% correct, respectively, whereas those for the posttest were 55% and 77%, respectively. In contrast, it is interesting that there is no significant interaction of instruction with either gender or grade. This means that the instructional gains were comparable for male and female students, for the reflective-assessment and control classes, and for all grades. They were also similar for all teachers at each grade level. Finally, although male and female students showed equivalent gains in physics knowledge, there was a difference in their average scores on the pretest of 7% (0.4 $\sigma$), which is significant, $F(1, 242) = 9.43, p = .002$. Male students had an average pretest score of 46%, whereas female students had an average score of 39%.

**Comparison with high school physics students.** To gauge how well the ThinkerTools students performed, we can compare their posttest performance with that of high school physics students taught using traditional approaches. The comparison data come from an earlier study in which we used a similar test of physics knowledge (B. White, 1993b). The set of six multiple-choice items that we have in common between these two studies cover topics addressed in the more difficult, later modules of the ThinkerTools curriculum: There are two items relating to each of Modules 4, 5, and 6 (i.e., Two-Dimensional Motion, Gravity, and Trajectories). Examples of the items are shown in Figures 22 and 23. They can all be characterized as simple questions that require the use of basic Newtonian principles to determine what will happen.

Both groups performed equally well on one of the items (85% correct), the ThinkerTools students outperformed the high school students on four items, and the high school students did better on one item. Overall, on these six items, the ThinkerTools students did significantly better, averaging 68% correct, whereas the high school students averaged 50% correct, $t(343) = 4.59, p < .001$. It is also

---

16This happens to be the only item without a diagram and thus relies more heavily than the other items on the students’ reading skills. Many of our middle school ThinkerTools students have poor reading skills, which may account for their weaker performance (50% vs. 83% correct) on this one item.
informative to compare the performance of the low-CTBS ThinkerTools students with that of the entire group of high school physics students. The low-CTBS ThinkerTools students averaged 58% correct, whereas the high school students averaged, as mentioned previously, 50% correct. Thus, even the low-achieving ThinkerTools students outperformed the high school physics students, \( t(167) = 1.97, p < .05 \).

It is well known that high school and college physics students frequently give incorrect answers to such problems, thereby exhibiting some of the well-documented misconceptions that have intrigued many researchers. Of course, the high school students would do better on standard physics problems found on high school and college tests—problems that can be solved via algebraic, constraint-based reasoning as opposed to qualitative, causal reasoning. We (and others) have argued that items like those shown in Figures 22 to 24 provide a better test of whether students understand the implications of Newton’s laws of motion (Clement, 1982; diSessa, 1982; McCloskey, 1983; McCloskey et al., 1980; McDermott, 1984; Minstrell, 1982; B. White, 1981, 1993b).

The success of the ThinkerTools students compared to the high school physics students is impressive because the high school physics classes represent a highly favorable situation for applying the traditional approach to physics education, namely, low class sizes, high-achieving students, and teachers with strong physics backgrounds. In contrast, two out of the three ThinkerTools teachers had little or no background in physics, and they are teaching in settings in which there are many low-achieving students and larger class sizes. Although the ThinkerTools teachers are working under more difficult conditions, their students are still able to outperform the high school physics students on these types of basic physics problems.

**Transfer.** We classified the items on the Applied Physics Test in terms of how much transfer they require from situations dealt with in the curriculum. An example of such a transfer item is shown in Figure 24. In this item, students need to reason about the initial velocity of the ball combined with the effects of gravitational attraction and how these affect the landing point of the ball. Although these factors have been explored by students within the curriculum, they have not been applied in the context of releasing an object while moving. Three levels of transfer were distinguished: no transfer (four items), near transfer (four items), and far transfer (two items). No-transfer items involve concepts and contexts that have already been encountered within the curriculum. Near-transfer items employ the same concepts but in slightly different contexts. Far-transfer items employ the same concepts but in totally new contexts.

In Table 5, we present the average percent correct on the pretest and posttest, along with the average gain, for each of the three levels of transfer. As one would expect, the posttest scores are highest for items that require the least transfer (the no-transfer and near-transfer items). However, the gain scores from pretest to
TABLE 5
Average Multiple-Choice Scores on the Applied Physics Test as a Function of Degree of Transfer Required

<table>
<thead>
<tr>
<th></th>
<th>Pretest (%)</th>
<th>Posttest (%)</th>
<th>Gain (%)</th>
<th>Effect Size</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>No transfer</td>
<td>52</td>
<td>76</td>
<td>24</td>
<td>1.2 σ</td>
<td>(1, 226) = 156.3</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Near transfer</td>
<td>33</td>
<td>63</td>
<td>30</td>
<td>1.1 σ</td>
<td>(1, 221) = 106.9</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Far transfer</td>
<td>47</td>
<td>57</td>
<td>10</td>
<td>0.3 σ</td>
<td>(1, 226) = 6.6</td>
<td>&lt; .01</td>
</tr>
</tbody>
</table>

The explanations score represents perhaps our most sensitive measure of applying physics knowledge because it is less susceptible to guessing and it provides an evaluation of the correctness and thoroughness of the students' written explanations for their answers. The results closely parallel those we found for the multiple-choice scores. Students showed large and significant improvements in their scores on their explanations, $F(1, 228) = 179.7, p < .001$, increasing from an average score of .50 on the pretest to .95 on the posttest (an average gain of .45). Measured in terms of the standard deviation of the pretest score distribution (.34), this represents an effect size of 1.3 σ. As was the case in the analysis of multiple-choice scores, there is a significant interaction of CTBS level with the effects of instruction, $F(1, 228) = 51.7, p < .001$. Average pretest scores for the low- and high-CTBS students were, respectively, .30 and .67, whereas corresponding averages for the posttest were .48 and 1.29, respectively. Measured in terms of effect sizes, the gains for high-CTBS students are equivalent for the multiple-choice (1.9 σ) and explanations scores (1.8 σ). However, for the low-CTBS students, the gains for the explanations scores (0.6 σ) were smaller than those for the multiple-choice scores (0.9 σ). This suggests that low-CTBS students have more difficulty in providing correct explanations for problems that they can otherwise solve correctly. Again, in contrast to these differences for low- and high-achieving students, we found that the instructional gains were the same for all grades, for both male and female students, and for the reflective assessment and control classes.

The explanations scoring rubric gives us an opportunity to judge the quality of the ThinkerTools students' performance against a very high standard based on a physicist's analysis of each problem. The average score of .95 on the posttest corresponds to a level on the scoring rubric for which an answer is correct but is
seen as providing an incomplete explanation—one in which they mention only a single variable and may have used physics concepts imprecisely. However, a clearer impression of the students’ posttest performance can be gained by looking at the percentage of students’ posttest explanations that were classified at each level of the scoring rubric. In all, 54% of students’ explanations on the posttest received a score of 0, indicating that they gave an incorrect answer. However, out of the 46% of explanations that were judged correct (i.e., that received a score of 1 or better), a greater percentage received a score of 2 (38%) or 3 (37%) than received a score of 1 (24%). Thus, when students were able to generate a correct explanation, it was most often an explanation that involved several variables and one that also accounted correctly for their causal interactions.

**Pretest differences among groups.** Although students who represented different grades or genders showed comparable gains, there were differences in their average scores on the pretest. The pretest explanations scores are a measure of students’ prior knowledge and preparation in science, particularly with regard to the physics of force and motion. In an analysis of pretest explanations scores, there are significant effects of CTBS, $F(1, 245) = 63.0, p < .001$; gender, $F(1, 245) = 33.8, p < .001$; and grade, $F(1, 245) = 5.9, p = .02$; and a marginally significant Gender $\times$ Grade interaction, $F(1, 245) = 2.9, p = .09$. With regard to gender and grade, in Grade 7, male and female students had average pretest explanations scores of .52 and .33, respectively, with a difference of .19 ($0.6 \sigma$). In Grades 8/9, this pretest difference grew to .34 ($1.0 \sigma$), with male students having a higher average score of .71 and female students having an average score of .37, which was nearly the same as that for seventh-grade female students. This trend is consistent with findings from an earlier study in which we compared pretest physics scores for sixth graders with those of high school physics students in Grades 11/12 (B. White, 1993b). The pattern of results was almost identical: The sixth-grade male students marginally outperformed the female students. However, the gender gap increased at the higher grades, so that the male students in Grades 11/12 significantly outperformed the male students in Grade 6, whereas the female students in Grades 11/12 were performing at the same level as the female students in Grade 6.

**Transfer.** The average explanations scores on the Applied Physics Test are given in Table 6 as a function of degree of transfer required. Again, the posttest scores are highest for items that require the least transfer (the no-transfer and near-transfer items), and the gain scores are significant for all of the three types of items. These results show that, when the criterion for successful performance is the ability to generate a correct and comprehensive explanation for one’s solution to a problem, the students still show evidence of transferring their conceptual model to novel, real-world situations.
TABLE 6
Average Explanations Scores on the Applied Physics Test as a Function of Degree of Transfer Required

<table>
<thead>
<tr>
<th>Effect</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Gain</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>No transfer</td>
<td>0.69</td>
<td>1.22</td>
<td>0.53</td>
<td>1.3 σ</td>
<td>(1, 228) = 135.3</td>
</tr>
<tr>
<td>Near transfer</td>
<td>0.27</td>
<td>0.88</td>
<td>0.61</td>
<td>1.1 σ</td>
<td>(1, 221) = 95.9</td>
</tr>
<tr>
<td>Far transfer</td>
<td>0.54</td>
<td>0.76</td>
<td>0.22</td>
<td>0.4 σ</td>
<td>(1, 227) = 9.8</td>
</tr>
</tbody>
</table>

Summary and Discussion of Physics Knowledge Results

Overall, the ThinkerTools Inquiry Curriculum was very effective in developing students’ conceptual models for force and motion and in enabling them to apply their models to analyzing new, real-world phenomena. Furthermore, their performance exceeded that of high school physics students (who were taught using standard physics curricula) on basic physics problems—problems that require the application of Newtonian principles to determine what will happen when objects are dropped, kicked, or hit.

With respect to the reflective-assessment treatment, the assessment process had a significant effect on the Conceptual Model Physics Test scores, particularly for the younger and lower achieving students. However, there was no evidence of any such effects on Applied Physics Test scores. Why should this be the case? The answer may lie (a) in the way the assessment concepts were defined and illustrated, as well as (b) in the context in which reflective assessment was practiced in the classroom. First, although the assessment concepts include aspects of performance such as “Reasoning Carefully,” which could in principle be applied to problem-solving tasks such as those in the physics tests, the Reflective-Assessment Process presented in the curriculum is directed primarily at students’ performance in carrying out research projects. For instance, “Reasoning Carefully” is defined for students using examples such as showing “whether or not a prediction or law that they or someone else has suggested fits with a scientific model” or showing “how experimental observations support or refute a model.” Second, students practiced reflective assessment primarily in the context of judging their own and others’ work on projects, not when they were solving problems. Thus, it is reasonable to expect that, although reflective assessment should have a direct effect on learning how to carry out inquiry, it should have only an indirect effect on the physics that was learned—through its influence on the learning of inquiry skills that (presumably) help students to learn physics principles. The effects of reflective assessment on the learning of physics should therefore be greatest for problems that are most closely related to the experiments that students carried out. This, we have seen, is the case: Reflective assessment improved performance on the Conceptual Models
Test, where problems are presented in a context similar to that of the computer simulation that the students used when they carried out their computer experiments, but not on the Applied Physics Test, where problems are presented in a context that often had little overlap with the situations students encountered in either their computer or real-world experiments.

Finally, we should bear in mind that reflective assessment is only one of a number of factors that contribute to students' success in learning how to do inquiry. Other factors include students' prior knowledge of inquiry, the composition of their research groups, and, most important, the support provided by the ThinkerTools Inquiry Curriculum itself. A more direct source of evidence for the effects of inquiry on physics learning lies in the correlations among measures of actual student learning outcomes in developing a knowledge of inquiry and of physics. These correlations are reported in the section to follow.

Relationship of Inquiry to Physics Learning

We have seen that participation in the ThinkerTools Inquiry Curriculum led to substantial improvements in students' understanding of Newtonian models for force and motion and in their ability to apply these models to real-world situations and explain their reasoning while doing so. These results lead to our second research question: What is the relationship between learning how to do inquiry and learning physics? In addressing this question, we are attempting to validate our contention that, within the ThinkerTools classes, students' inquiry activities are instrumental in enabling them to construct their conceptual models of force and motion.

It is clear that students' learning of physics is determined by multiple factors. These include the students' use of inquiry processes in creating and verifying physical laws and models, but are also likely to include students' prior knowledge of physical science and their learning of physics through direct instruction by teachers or via the curricular and computer materials. To evaluate the separate influence of inquiry on physics learning, we can compare posttest measures of the physics that students have learned with pretest measures of their physics knowledge and of inquiry, as well as with measures of inquiry derived from their project work within the curriculum and their inquiry posttests. Evidence for the direct role of inquiry processes in students' learning of physics principles would consist of significant correlations of physics learning with inquiry measures, and these should remain high when we control statistically for students' degree of prior physics knowledge.

In Table 7, we present the correlations between three posttest assessments of students' physics and inquiry performance with pretest measures of their knowledge of physics and inquiry, as well as with several measures of the inquiry skills they acquired from their work in the class. The learning outcome measures include
TABLE 7
Correlations of Final Assessments of Learning With Pretests and Inquiry Assessments

<table>
<thead>
<tr>
<th>Learning Measures</th>
<th>Pretests</th>
<th>Inquiry Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applied</td>
<td>Inquiry</td>
</tr>
<tr>
<td></td>
<td>Physics</td>
<td>Pretest</td>
</tr>
<tr>
<td></td>
<td>Pretest</td>
<td>Score</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>Score</td>
</tr>
<tr>
<td>Conceptual Physics</td>
<td>.19* (.29)</td>
<td>.38* (.40)</td>
</tr>
<tr>
<td>posttest</td>
<td></td>
<td>.32* (.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.27* (.34)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.50* (.54)</td>
</tr>
<tr>
<td>Applied Physics</td>
<td>.31* (.47)</td>
<td>.43* (.46)</td>
</tr>
<tr>
<td>posttest</td>
<td></td>
<td>.24* (.30)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.22* (.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.47* (.51)</td>
</tr>
<tr>
<td>Inquiry posttest</td>
<td>.10 (.15)</td>
<td>.56* (.60)</td>
</tr>
<tr>
<td></td>
<td>.44* (.53)</td>
<td>.38* (.48)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
</tr>
</tbody>
</table>

Note. Correlations in parentheses are corrected for unreliability of the column measures. The reliabilities used in making these corrections are .42 for Applied Physics pretests, .88 for Inquiry pretests, .69 for Mass Project overall scores, .62 for Final Project overall scores, and .86 for Inquiry posttests.

*p < .01.

(a) the students' scores on the Conceptual Physics Test, (b) the students’ scores on the Applied Physics Test, and (c) the students’ posttest scores on the Inquiry Test. Measures of students’ initial knowledge of physics and of inquiry include (a) their pretest scores on the Applied Physics Test and (b) their pretest scores on the Inquiry Test. Measures of students’ acquired inquiry skill and knowledge are (a) their overall scores on their Mass Project, (b) their overall scores on their Final Project, and (c) their posttest scores on the Inquiry Test.

The first interesting finding is that measures of students’ acquired physics knowledge correlate as highly or more highly with their prior knowledge of inquiry than with their prior knowledge of physics. For Conceptual Physics Test scores, the correlation with the Inquiry Pretest scores is .38, whereas the correlation with the Applied Physics Pretest is .19; the difference between these correlations is marginally significant, t(107) = 1.51, p = .07. Similarly, for Applied Physics Posttest scores, the corresponding correlations are .43 with the Inquiry Pretest and .31 with the Applied Physics Pretest.17

Second, we find that students’ physics posttest performance is also strongly related to their developed knowledge of inquiry, shown by their performance on inquiry tasks completed at the end of instruction. Posttest scores on the Conceptual

---

17 This and the remaining findings presented in this section depend on the relative sizes of various correlations shown in Table 7. The possibility exists that two correlations can differ due to differences in the reliabilities of the measures being correlated. To check for this possibility, we recalculated all of the correlations, correcting them for the unreliabilities of the measures. These corrected correlations are also shown in Table 7 (in parentheses). Although the absolute sizes of the corrected correlations are often appreciably larger, the relative sizes of the correlations discussed in this section are not substantially changed.
Physics Test correlate .50 with the Inquiry Posttest scores, and Applied Physics scores have a correlation of .47. Physics posttest scores also correlate significantly with scores on inquiry projects students have undertaken during the curriculum (their Mass Projects and Final Projects). The correlations between students' Conceptual Physics scores and their project scores are .32 for Mass Projects and .27 for Final Projects. The corresponding correlations for Applied Physics scores are .24 and .22, respectively. We should note that the students' project scores were based on write-ups of group work, so it is to be expected that these correlations are lower than the correlations with the Inquiry posttests, which were taken by each student individually.

Third, the correlations between measures of students' acquired physics knowledge and their acquired skill in inquiry are not reduced when we control statistically for students' prior knowledge of physics using partial correlation analysis. For example, the partial correlation of Conceptual Physics scores with Inquiry posttest scores controlling for Applied Physics pretest scores is .49 (compared with a simple correlation of .50), and the corresponding partial correlation for Applied Physics posttest scores is .46 (compared with a simple correlation of .47). Thus, we can conclude that the knowledge of scientific inquiry that students develop in the course of the curriculum has a strong effect on students' learning of physics, and this is independent of their prior knowledge of physics.

Finally, although students' acquired knowledge of physics correlates positively with their prior knowledge of inquiry (their Applied Physics Posttest scores correlate .43 with their Inquiry Pretest scores), the reverse is not true: Students' acquired knowledge of inquiry is not related to their prior knowledge of physics (their Inquiry Posttest scores correlate only .10 with their Applied Physics Pretest scores). At the same time, their Inquiry Posttest scores correlate .56 with their Inquiry Pretest scores. Thus, although the learning of physics is strongly related to the quality of students' prior inquiry skills, the reverse is not true—their learning of inquiry is unrelated to their prior knowledge of science—although it is related to their prior inquiry knowledge.

Taken together, these results show that inquiry skills play a significant role within ThinkerTools classes in the students' learning of physics. This validates our hypothesis that students are using their skill in inquiry to develop an understanding of the physics, as should be the case if their inquiry is serving its intended purpose of helping them to construct models for the phenomena they are observing. However, there are other factors that also come into play in learning the physical principles. It is, after all, possible for students to do good inquiry and still discover incorrect physics because real-world experimentation is prone to observational errors and errors of measurement. It is our belief that the computer simulations and the experimental activities using the computer are a major factor in helping students develop correct physics laws. Our prior research indicates that interacting with these computer simulations without going through the inquiry process can produce
significant gains in physics test scores (B. White, 1981, 1984). However, this observation does not diminish the importance of combining the representational and observational clarity of the simulation with well-thought-out inquiry processes so that students can participate in the process of generating models and theories in science.

RESULTS: ATTITUDES ABOUT LEARNING SCIENCE

Students’ beliefs about the nature of scientific knowledge and the processes of learning science are influenced for better or for worse by the various science curricula they take in school (Carey & Smith, 1995; Driver, Leach, Millar, & Scott, 1996; Hammer, 1994; Schomer, 1993; Songer & Linn, 1991). To investigate how such beliefs were influenced by the ThinkerTools curriculum, the students were given a questionnaire before and after the curriculum (some of the items were taken from questionnaires developed by Aikenhead & Ryan, 1992; Linn & Songer, 1993; Ryan & Aikenhead, 1992; Songer & Linn, 1991). The results indicate pretest to posttest changes in their attitudes and beliefs about learning and doing science.

Results. On this questionnaire, which included 30 attitudinal-scale items, there were significant or marginally significant changes in the students’ responses to 15 of the items. We present here representative examples of students’ responses. Students showed a significant or marginally significant increase in their disagreement with the following statements in their posttests as compared with their pretests:

1. Views about scientific theories:

   • “The theories of science have always been true and will always be true,”18 \( t(243) = 1.73, p = .04 \), with disagreements increasing from 70% to 73% and agreements decreasing from 10% to 8%.

2. Views about learning science:

   • “A lot of things in science must simply be accepted as true and remembered; there aren’t explanations for them,” \( t(223) = 3.52, p = .001 \), with disagreements increasing from 51% to 62% and agreements decreasing from 21% to 16%.

18For the following items, students circled a rating on a 5-point scale ranging from 1 (strongly disagree) through 3 (neutral) to 5 (strongly agree). Tests of significance are t tests of pretest to posttest differences in mean ratings. Percentage agreement refers to a choice of scale values 4 or 5; disagreement refers to a choice of scale values 1 or 2.
• “A lot of science is mathematical and abstract. There just isn’t a way to understand it intuitively,” \( t(247) = 1.41, p = .08 \), with disagreements increasing from 39% to 45% and agreements decreasing from 24% to 19%.

• “When learning new science material, I prefer to be told what is correct by a teacher,” \( z = 4.48, p < .001 \). Choices of always decreased from 29% to 18%; sometimes or never increased from 71% to 82%.

• “When learning new science material, I prefer to read the right answers in the textbook,” \( z = 2.57, p = .005 \). Choices of always decreased from 13% to 8%; sometimes or never increased from 87% to 92%.

3. Views about scientific “aptitude”:

• “To be good at science, you need to have a kind of ‘scientific mind,’” \( t(248) = 2.56, p = .005 \), with disagreements increasing from 41% to 50% and agreements decreasing from 26% to 22%.

• “In general, boys tend to be naturally better at science than girls,” \( t(247) = 1.38, p = .08 \), with disagreements remaining at 83% and agreements decreasing from 8% to 4%.

Summary and Discussion of Attitudinal Results

These results indicate that, with the ThinkerTools Inquiry Curriculum, students are less likely to view scientific theories as immutable and never subject to revision. They tend to see science as more meaningful and explainable, rather than as abstract and purely mathematical. They also prefer to take a more active role in learning science, rather than simply being told what is correct. The students’ attitudes about scientific aptitude also change. Students are less likely to attribute being good at science to an inborn ability—having a “scientific mind”—or to agree with a gender stereotype about scientific aptitude. Interestingly, the largest changes in attitudes about scientific aptitude (shown in the “scientific mind” question) are for female students and for students with low CTBS scores.\(^{20}\) And the largest changes in gender

\(^{19}\)For the following two items, students circled one of three categories: always, sometimes, or never. Tests of significance are \( z \) tests of pretest to posttest differences in the proportion of students choosing always.

\(^{20}\)For female students, there is a significant change in mean ratings, \( t(114) = 2.55, p = .006 \), with disagreements increasing from 43% to 57% and agreements decreasing from 21% to 13%. For male students, there is no significant change in mean rating, \( t(106) = 1.10, p = .14 \), with disagreements increasing from 40% to 43% and agreements decreasing from 32% to 31%.

With regard to achievement level, three levels were used in the analysis of the attitudinal measures: low, middle, and high. For low- and middle-CTBS students (i.e., those with CTBS scores lower than the 80th percentile), the change in mean ratings is significant, \( t(140) = 3.64, p < .001 \), with disagreements increasing from 38% to 52% and agreements decreasing from 30% to 19%. There are no significant changes for high-CTBS students, with disagreements increasing from 49% to 54% and agreements remaining at 18%. Thus, the low- and middle-CTBS students’ attitudes have become equivalent to those of the high-CTBS students.
attribution of scientific aptitude (shown in the "boys are better than girls" question) are for low-CTBS students. These attitudinal changes, especially those of the lower achieving students and female students, are encouraging. In general, the findings suggest that the curriculum is effective not only in improving students' understanding of inquiry and physics but also in improving their views about the nature of scientific knowledge and learning.

**GENERAL DISCUSSION**

**Implications of Our Findings**

In this section we summarize our major findings and discuss their implications for inquiry-based science education and for reforming assessment practices.

Our first major finding is that one can successfully teach sophisticated, inquiry-based science in urban schools. We found that research-based science classes create a climate in which students of widely varying backgrounds and levels of academic achievement can be actively engaged in their research and participate fully in the classroom community. In such classes, we found significant pretest to posttest gains on a physics test for all groups of students, regardless of their achievement levels. Although the low-CTBS students did not do as well as the high-CTBS students, they still outperformed high school physics students on qualitative problems in which they were asked to apply the basic principles of Newtonian mechanics to real-world situations. Similarly, for students in the reflective-assessment classes, there were significant pretest to posttest gains for all groups of students on a test of scientific inquiry skills, regardless of their achievement levels. In fact, the gain scores were higher for the lower achieving students. On this test, the largest improvements were for the measure of Coherence—the measure that captures the logical relations among hypotheses, experimental design, and the interpretation of results. In addition, low-CTBS students who submitted write-ups of their research projects were able to approach the level of high-CTBS students in the quality of their submissions—particularly if they worked collaboratively with high-achieving students in reflective-assessment classes. Inquiry thus appears to be accessible to a wide range of students with varying prior educational advantages and accomplishments.

Such a curriculum introduces students to an inquiry process that is authentic (it is applicable to doctoral theses as well as middle school research projects) and that

---

21For low-CTBS students (i.e., those with CTBS scores lower than the 45th percentile), there is a significant change in mean ratings, t(53) = 1.82, p < .04, with disagreements increasing from 74% to 80% and agreements decreasing from 17% to only 4%. For middle- and high-CTBS students, there is no significant change in mean rating, t(156) = -0.10, with disagreements decreasing from 87% to 83% and agreements remaining at 5%. Again, on the posttests, the attitudes of the low-CTBS students are similar to those of the middle- and high-CTBS students.
incorporates a modern view of the epistemology of science (cf. Chalmers, 1990; Collins & Ferguson, 1993; Matthews, 1994; Nadeau & Desautels, 1984; Nersessian, 1992). It views theories as constructed by scientists and recognizes that the observational base for the theory is itself derived from theory. Furthermore, the objectivity of results is derived from agreement within the scientific community (intersubjectivity) in the concepts and theoretical derivations used in establishing an observational base and in making theoretical inferences based on it.

Our students participated in a classroom research community that embodied each of these ideas within its structure. Thus, students generated research questions and conjectures on the basis of their prior theories (both informal and formal), they negotiated among themselves to decide what observations and experiments would yield useful information to test their conjectures, and they developed laws to formally describe their findings in a form that enabled them to generalize (apply) their results to new situations. Our view is that the best way for students to understand the epistemology of science is to engage in practices that involve that epistemology, rather that being told about it. The knowledge they develop from reflecting on their own practices as a scientific community should enable them to understand better the history and philosophy of science as well as to evaluate public policy debates where models and results of science play a role (e.g., effects of human activity on global warming). Further research is needed, of course, to evaluate this conjecture.

We also found that learning inquiry improves students’ learning of science concepts, laws, and models as well as their ability to use them in analyzing new situations. We found correlations between measures of students’ success in learning and applying science knowledge and their success in learning to do inquiry. Within the ThinkerTools Inquiry Curriculum, learning science is a constructive act—students individually and as a group build their knowledge of science concepts, laws, and models. Our conjecture is that science knowledge developed in this way is better understood by students and more useful to them than when it is directly taught (cf. A. Brown & Campione, 1996; Carey, Evans, Honda, Jay, & Unger, 1989; Champagne, Klopfner, & Anderson, 1980; Collins & Brown, 1988; diSessa, Hammer, Sherin, & Kolapakowski, 1991; Driver, Asika, Leach, Mortimer, & Scott, 1994; Duschl & Gitomer, 1997; Hatano & Inagaki, 1991; Ministrell, 1989; Scardamalia & Bereiter, 1991). Our comparison of urban, middle school students’ qualitative understanding and application of science knowledge with that of high school physics students provides evidence for this conclusion. Classroom practices that have contributed to understanding and using inquiry include an emphasis on having students develop experiments and laws and on having frequent opportunities for explaining their experiments and laws to others (Chi et al., 1989). In the ThinkerTools classes, explanation played a critical role in the classroom research process; indeed, the acceptance of
a suggested law or model within the classroom research community depends on presenting it clearly and showing how it is supported by observation and experiment. Likewise, the ability to apply models and laws to make predictions for new situations is recognized within the inquiry process as an aspect of verification and as a means for extending the generality of one’s model.

Inquiry-based science employed in the ThinkerTools Curriculum develops the important link between inquiry practices and scientific knowledge, a link that is poorly developed within conventional science curricula. Conventional curricula are often incoherent in their treatment of inquiry: A unit on “the scientific method” is often followed by classroom practices that do not employ inquiry processes to construct scientific knowledge (i.e., concepts and models). We have found that there is little correlation between students’ science content knowledge and their knowledge and use of inquiry (the correlation between pretest scores is .03) in a population of students that has received the conventional curriculum in their prior schooling. Our conjecture is that this lack of correlation is due to there being no instructional or functional linkage between inquiry processes and the learning of science in conventional science instruction. Knowledge of science and of inquiry are disconnected. Yet, in the ThinkerTools Inquiry Curriculum, we have demonstrated such a link, showing that inquiry can become a basis for learning science content knowledge.

Learning inquiry is particularly effective in meeting the needs of educationally disadvantaged students. We found that there are large pretest deficits in knowledge of inquiry processes among less-advantaged students (students with low CTBS scores). Our conjecture is that this is due to their having had little exposure to any clear, understandable process for inquiry, unlike advantaged students who have had exposure to inquiry activities that involve generating plausible theories and explanations for various situations as well as evaluating and revising their theories based on observations and theory-driven arguments. We found that the ThinkerTools Inquiry Curriculum, particularly when combined with reflective-assessment practices and heterogeneous grouping of students, is very effective in reducing performance gaps between educationally disadvantaged and advantaged students in their understanding and use of the inquiry process. We believe this is due in part to the initial introduction of the Inquiry Cycle as a set of ordered activities that students can easily follow. A student’s deeper understanding of the significance of these activities comes from actually practicing them and reflecting on them over time. The hypothesis is that one has to be already engaged in a practice in order to develop an understanding of it, because if you are not doing it, you cannot reflect on it. This theory of learning by doing and reflecting also underlies other successful educational practices such as using Reciprocal Teaching to develop skills and expertise in reading comprehension (A. Brown & Palincsar,
1989; Palincsar & Brown, 1984). So, our general conclusion is that inquiry can and should be taught as early as possible using an inquiry-oriented approach like that of the ThinkerTools Inquiry Curriculum, in which low-achieving students work in partnership with higher achieving students to plan, carry out, and critically evaluate research. Such an approach develops students’ inquiry and reflective processes, which are important to scientific expertise and to learning in general.

Introducing inquiry-oriented curricula and assessments is important if we are to address gender differences in learning science. We found that there are no gender differences in the learning of scientific inquiry. At the same time, we found that there are gender differences in science knowledge. These differences appeared in our pretest measures of physics knowledge. And, although we found equal gains in physics test scores for both male and female students, there were still posttest differences at the end of the curriculum.

These findings have important implications for gender equity in the assessment of students’ learning about science. The argument goes as follows: First, most curricular frameworks now recognize the importance of students’ understanding the epistemology of science and how to carry out inquiry processes. Indeed, inquiry knowledge and skills are arguably important predictors of future accomplishments in science. Furthermore, we have seen that learning of inquiry improves learning of science content knowledge within an inquiry-oriented curriculum. Second, if scientific inquiry knowledge is valued and useful in learning science, then it should be an important component in assessing science learning and accomplishment. We have demonstrated in our work a number of ways to reliably assess students’ ability to carry out science inquiry, using both paper-and-pencil tests and research projects. Third, in these assessments we have found that, in contrast to science content knowledge, there are no gender differences in performance on inquiry assessments, either prior to or following instruction. Therefore, because it is possible to reliably assess such inquiry processes and because such assessments are free of gender differences, then to ensure a full and fair appraisal of male and female students’ understanding of science, science assessments must be broadened to include such inquiry processes. Moreover, because assessment can drive curriculum (Madaus, 1988), valuing inquiry in science assessments will be an important factor in encouraging the introduction of inquiry-oriented science instruction into school curricula.

It is desirable to introduce inquiry-based science early in the school curriculum. Another major implication of our research is that inquiry and reflective assessment should be taught early. This would enable young students to develop inquiry and metacognitive skills that are important components of expertise in learning. These skills should help low-achieving students to overcome their educational disadvantages.
This conclusion is based on two results. First, students over a range of grades showed equal degrees of learning using the inquiry curriculum: We found no age differences in students' pretest or posttest scores on the inquiry test from Grade 7 to Grade 9, and we did not find any age differences in students' gains on the physics tests. Moreover, our earlier research shows that sixth-grade students can also do as well on comparable problems (B. White, 1993b). So, from the standpoint of readiness to learn, the ThinkerTools Curriculum is as appropriate for sixth- and seventh-grade students as it is for eighth and ninth graders. Second, in our pretests we found evidence of gender differences in science content knowledge. The interesting finding is that there were only small differences between male and female students in the seventh grade, but these gender differences grew larger in the later grades. In an earlier study, comparing the performance of sixth graders with high school students in Grades 11/12, we found a similar pattern: There were marginal differences at the sixth-grade level and a significant increase in the gender gap by the high school level (B. White, 1993b).

Together, these results suggest that inquiry-based science could and should be introduced in earlier grades (see also Metz, 1995) and that doing so in these grades (before gender differences in science content knowledge have developed) might help to eliminate the gender differences that develop in knowledge of and interest in science. The subject of force and motion is appropriate for elementary school students, who are involved in sports and focused on the physical world in general. Also, early introduction of an inquiry curriculum would enable students to learn inquiry and metacognitive skills that are not only helpful in learning science but are transferable to other subjects as well. These skills should help the low-achieving students to overcome educational disadvantages. Furthermore, teaching inquiry may also improve students' attitudes toward learning and doing science.

Based on these findings, our participating schools are now introducing ThinkerTools at the sixth- and seventh-grade level. In later grades, students go on to a project-oriented curriculum in which they use their understanding of inquiry to do research projects. Some of their research projects involve physics, but some do not. Many students are interested in other topics such as educational research, and they investigate questions like whether “playing music helps you study” or whether “females who do research with female partners as opposed to male partners do better.” Investigating such questions raises many interesting issues and enables them to do some stimulating research projects. In general, we argue that inquiry knowledge and skills should be taught early so that, by the higher grades, students can be choosing research topics that are of particular interest to them. These may be outside of the standard science curriculum topics (i.e., chemistry, physics, biology, and earth sciences) and can include research questions that relate to education, psychology, and sociology, which are particularly interesting for many students at the middle school level.
Inquiry-oriented science and the ThinkerTools Inquiry Curriculum in particular appear to change students' views of aptitude for learning and understanding science. We found that there were significant decreases in rates at which students agreed with the view that “to be good in science, you need to have a kind of ‘scientific mind’” and with conceptions of science such as “a lot of things in science must be accepted as true and remembered; there aren’t explanations for them.” Viewing science learning as determined by one’s aptitude is a damaging conception, especially when one is attempting to learn material that is perceived as difficult (Dweck & Bempechat, 1983). This may be particularly the case for lower achieving students and for female students, who are more likely to underestimate their skill and overestimate the difficulty of a task (Dweck, Davidson, Nelson, & Enna, 1978). In fact, aptitude concepts are endemic in our culture, and they can drive out the alternative of viewing science knowledge as the result of concentrated attention to learning over an extended period of time. Our results show that making inquiry processes explicit through the Inquiry Cycle and applying them in the context of actually practicing inquiry with one’s peers demystifies what scientific inquiry is and allows students to see that they too can do science. This combination of valuing inquiry and showing how it is done allows students to see that it is within their grasp, rather than limited to a naturally endowed scientific elite.

It is possible for classroom teachers to reliably assess students’ inquiry skills, using both paper-and-pencil tests and research projects, and assessing such skills is important if we are to avoid bias against both lower achieving students and female students. In a separate article (Frederiksen & White, 1997b), we presented a detailed study of teachers scoring inquiry projects. Our findings are that classroom teachers were able to make reliable judgments as they scored students’ Common Projects as well as their Chosen Projects. Furthermore, we believe that learning how to use the scoring categories to assess their students’ research helps teachers understand the inquiry process and makes them more effective in facilitating students’ work. Having students also learn how to apply the scoring categories and to reflect on their work makes it easier for teachers to have conversations with students about their performance. For these reasons, it is important to incorporate measures of inquiry, such as our inquiry tests and research projects, when assessing the effects of a science curriculum.

Another reason for broadening the range of science assessments to include inquiry measures is that if only physics tests are used, the results tend to be more biased against both low-achieving students and female students. For instance, on the inquiry tests and research projects, we found that low-achieving students who had the benefit of the reflective-assessment classes did almost as well as the high-achieving students. Furthermore, these results could not be attributed simply
to ceiling effects. We also found that the male and female students did equally well on the inquiry tests and projects. However, on the physics tests, the pattern of results was not comparable: Male students outperformed female students (on both pretests and posttests), and the high-achieving students outperformed the low-achieving students. Thus, utilizing inquiry tests and research projects in addition to subject matter tests not only plays a valuable role in facilitating the development of inquiry skills, it also produces a more comprehensive and equitable assessment of students’ accomplishments in learning science.

Reflective assessment provides an explicit classroom activity that brings metacognition into the social processes of the classroom, which enhances the acquisition of metacognitive knowledge and skills. Learning to monitor the quality of one’s thought and the products of one’s effort is the hallmark of what is meant by metacognition (Brown et al., 1983). We have found that reflective assessment, explicitly taught and practiced within the ThinkerTools Inquiry Curriculum, greatly improves the learning of inquiry for disadvantaged students as well as being beneficial for all students. Self-assessment is one way of introducing self-monitoring and evaluating into students’ habits of thinking and work. Our strategy has been to begin by choosing assessment concepts that represent particular aspects of metacognition that complement one another and that are applicable to a wide range of cognitive contexts (Frederiksen, 1994; B. White & Frederiksen, 1994). The following are some examples:

- **Inventiveness**: Using divergent thinking in generating ideas, finding questions, generating hypotheses, designing experiments, and applying models.
- **Systematicity**: Creating and using plans and monitoring their success as they are applied, for example, in conducting experiments.
- **Reasoning**: Developing logical arguments that are consistent in using laws and principles (perhaps of one’s own invention) to make predictions, design experiments, and explain findings.

The implicit overall goal is teaching about thought by providing conceptual categories that permit one to reflect on and talk about one’s thought processes in carrying out scientific inquiry. A second goal is to create an activity context (here it is assessment) in which metacognitive reflection can take place within the classroom. We have found that reflective assessment appears to be acceptable and enjoyable (to a point) to students. Furthermore, our analyses show that the effects of reflective assessment depend on the students’ understanding of the assessment concepts and their ability to use them in evaluating the quality of their work.

There are two important caveats that need to be understood if one is to introduce reflective-assessment practices into the classroom:
1. All participants must understand that it is performance that is being rated, not people, where performance is what you actually do, not what you are capable of doing.

2. Students must be given the means to understand how to do well in their performances; otherwise, performance ratings may be damaging to students.

These caveats relate to a very serious issue: Reflective assessment can be seen as a performance evaluation if it occurs in an instructional context that teaches students how to carry out good inquiry projects. However, if the process of producing a high-quality project is mysterious to students, they are likely to fall back on an ability attribution for the assessment results, that is, on the belief that they are not "smart enough" to do well in science. There is a clear equity issue here as well, because failure to provide both an understanding of the assessment criteria and of how to perform well may be particularly damaging to less-advantaged students who, without a clear understanding of how highly rated work is produced, are likely to invoke the damaging theory of performance as a reflection of their ability. So, reflective assessment should not be added on to a curriculum, rather it should be an integral part of a curriculum that scaffolds the development of the skills being assessed.

Middle school science teachers can successfully make use of the ThinkerTools Inquiry Curriculum and Assessments when we adopt the same constructivist approach to teachers' professional development that we do to students' learning. Teaching inquiry-based science can be considered a tall order for middle school teachers who have themselves learned science through conventional curricula and who are unlikely to have engaged in authentic scientific inquiry themselves in the course of their undergraduate education. The teachers may thus be in a similar position to their students in that they may also be encountering inquiry-based science education for the first time. We therefore take a similar approach to teachers' professional development that we take to students' learning scientific inquiry processes. The idea is to scaffold carefully the initial implementation of the curriculum so that teachers can get started doing it and then to encourage collaborative reflection on their performance with others who are so engaged, such as by sharing videotapes of classes or examples of students' work.

Our results suggest that physics and inquiry can be effectively taught even though the teachers may not initially have the subject matter or research expertise. Working with the ThinkerTools Curriculum—which uses computer models that embody the correct physics combined with the Inquiry Cycle, the scaffolded inquiry activities, and the Reflective-Assessment Process—can enable teachers with no initial background in physics or inquiry to get impressive results, both with regard to the students' inquiry skills and their physics understanding. Consider the results for one of our teachers, who was not involved in the development of the curriculum
and who did not have any formal physics background prior to implementing the curriculum. Her students performed as well as those of the other teachers on the physics tests, the inquiry test, and the projects. In fact, the results are, for the most part, the same for all of our teachers despite the fact that they have quite different teaching styles and different initial knowledge of physics. We are therefore optimistic that the combination of scaffolding the initial implementation of the curriculum (through the teacher’s guides and student research books) and providing opportunities to reflect on how it is going with other teachers is a good way to introduce teachers to new approaches to science education. Such an approach has the advantage of allowing dissemination of teaching practices to take place within the school environment with teachers helping other teachers, rather than through expensive workshops and centralized delivery of professional development services. We are currently conducting a larger scale study of the problems involved in disseminating the ThinkerTools Inquiry Curriculum and the effects of teachers’ implementing such an inquiry-oriented approach on their teaching practices in non-ThinkerTools portions of their curriculum.

Problems Arising From the Limitations of Our Work

We do not want to give the erroneous impression that the ThinkerTools curriculum produces perfect results and that there have not been any struggles in its creation. We will now discuss some of the challenges we faced.

Problems in the Development of Physics Expertise

One challenge relates to the students’ development of a Newtonian conceptual model of force and motion. The students’ written explanations to test questions and their verbal explanations given in the interviews conducted at the end of the curriculum indicate that some students who gave correct answers to physics test questions still showed evidence of common misconceptions. In one particularly prevalent type of misconception, often called impetus theory, students believe that, when you kick or throw an object, you give it a force that the object carries with it and that causes it to move. An example is the following quote from a student who was asked to explain what happens when you throw a ball up into the air:

When your hand throws the ball upward, it gives a force to the ball, so the ball is propelled upward. By the time the force wears out, then gravity—well, gravity is working too. It slows the ball down as it goes up. So, by the time the ball loses its force, gravity pulls it back down again.

This is an interesting answer because it incorporates both correct and incorrect physics: The student has the correct conception that gravity acts from the moment
the ball leaves your hand, but he also has the misconception that the ball carries an upward force that wears out. Such an interplay of correct and incorrect Newtonian reasoning was common in students’ answers and in some of the teachers’ answers to interview questions, as well as being present in Newton’s own thinking for many years (Steinberg et al., 1990). It may well reflect a reasonable transitional state on the path to expertise.

In subsequent versions of the curriculum, we introduced additional “challenge questions” aimed at provoking students to derive more sophisticated and accurate versions of Newton’s first and third laws of motion and to thereby overcome such impetus misconceptions. In pursuing these questions, the hope is that students will argue and come to agree that a force only occurs when two things interact and that an object does not need to have an internal or external force to keep it moving. For instance, to stimulate students to realize that, when objects interact, there is an equal and opposite action and reaction, we asked them to think about what happens when a bonker hits a ball. Creating inquiry activities that stimulate students to develop a Newtonian theory of what forces are, how they act, and when they act is nontrivial. It needs to be done so that it presents students with challenging thought experiments and research questions without taking too much of the initiative away from the students. The curriculum could use further revision in this regard, especially with respect to stimulating an understanding of what gravity is and how it works. This is a difficult research question with which theoretical physicists are still struggling. The students find questions related to mass and gravity highly interesting, but many of their erroneous theories survive the present version of the curriculum (Schwarz, 1995).

Problems in the Development of Inquiry Expertise

With regard to the students’ understanding of modeling and inquiry, there are some interesting challenges. We asked students the following question in an interview at the end of the curriculum: “If you did a computer experiment and a real-world experiment and you got different results, which one would you believe and why?” We found that students have intriguingly different theories about the relationship among the computer model, their own conceptual model, and the real world. For example, some students, roughly one third of those interviewed, believe that the computer model embodies the right physics, and so their job is to discover its laws in order to create their conceptual model and to make sure that their real-world experiments come out with the correct results. This is not an unreasonable view and suggests that these students understand the basic idea of a model.

Other students appear to lack such a modeling perspective and make statements such as the following: “The computer is just a machine and so it cannot hit a ball or know any physics. What I really believe is the results of my real-world experiments because that is the only thing that is real.” These students do not
appreciate that the computer simulation is following laws embedded in a reasoning structure (i.e., stepping through time and using laws to calculate changes in velocity based on the forces that are acting) and that versions of these laws and reasoning structures are needed for their conceptual model. Introducing them to a more explicit representation of how the computer simulation works may develop this understanding. For example, in some of our earlier research we created simulations that could, at the students’ request, illustrate such model-based reasoning by thinking out loud (B. White & Frederiksen, 1990).

Some students, in contrast, did appear to have the desired modeling perspective. They explained that the computer simulation is only a model and is an accurate model of the real world only if you include the right parameters. For example, you need to put in air resistance when there is significant air resistance, and you need to choose the correct law for resistance such as choosing gas–fluid resistance and not sliding resistance when appropriate. So, if you put in the right parameters and choose the correct law for the computer model, then the computer model will be an accurate model of the real world.

To help students develop the desired modeling perspective, we introduced a variety of activities such as having them discuss the results of their computer and real-world experiments to talk about discrepancies and their possible origins. Also, we included an interesting activity in which students watch a Wile E. Coyote and Road Runner cartoon and try to determine whether the laws of physics in the cartoon world are the same for Wile E. Coyote as they are for Road Runner, and whether any of the cartoon laws are different from those that describe how the real-world behaves. Discussions like these help students understand and think about the nature of laws and models in general.

Subsequently, we have also modified the ThinkerTools software so that students can easily create alternative models which embody competing theories of force and motion. The goal is to widen the range of inquiry strategies that are made possible by the software, as well as to provide students with a better tool for learning about the nature of scientific models (B. White & Schwarz, in press). To illustrate, students can now create, for example, a world in which objects spontaneously slow down, even when no forces are acting, or in which gravity causes objects to fall at a constant speed, even when there is no air resistance. By choosing from a set of alternative laws, students can construct worlds that obey the laws they select. Also, the software can now talk so that students can go into “step-through-time mode” and hear how the computer model is using the law they selected to determine what will happen. Students can thus use the software to reify and see the implications of their alternative hypotheses. They can also use it to embody the laws and conceptual models that they

---

22In the cartoon, the “physical laws” differ substantially for the two characters, which is a major source of humor in the cartoon.
derive from their real-world experiments and can compare these with Newtonian models. Schwarz created a version of the ThinkerTools Inquiry Curriculum that uses these new software features and inquiry strategies, and is investigating whether they enable students to better understand the nature of scientific models and the inquiry processes needed to create them (Schwarz, White, & Frederiksen, 1997).

Problems in the Development of Reflective-Assessment Expertise

We also found, in the early trials of this curriculum, that the students were reluctant to criticize each other’s research. For example, one group of students who did mass and gravity projects gave a presentation in which they concluded that heavier things fall faster. Then, the next group gave a presentation in which they concluded that they fall at the same rate, and no one pointed out the discrepancies in these results.

To encourage students to be more critical and reflective about their research, we added several activities to the curriculum. In one such activity, the students criticize anonymous research reports (as shown in Appendix A). These reports provide students with a model of how to write a research report. Their anonymity also enables students to feel comfortable in criticizing the research and in suggesting improvements. Another example of such an activity was developed by one of the teachers for use when the class is trying to reach a consensus about the law that best characterizes the findings from their experiments. The teacher has the students write their data on the board in a large table. The class then goes through each data set and formulates a law to characterize it. The students thereby see which data sets yield results that are discrepant with the most common findings, and they discuss why such discrepancies may have occurred. The class concludes by voting as to which law they think is the most accurate and useful, and if there are still major disagreements, the teacher has students prepare for and engage in a formal debate.

The teachers report that these activities, like criticizing anonymous reports and collaboratively analyzing data, are effective in activating and honing the students’ ability to reflect on research designs, analyses, and reports. They also serve to create a social environment within the classroom that more fully approximates that of a community of researchers.

Another challenge, which is related to reflective assessment, concerns the use of the scoring criteria (shown in Figure 8—such as “Using the Tools of Science” and “Reasoning Carefully”). The students used these criteria for peer and self-assessment, as described earlier. This Reflective-Assessment Process would probably have been even more effective and authentic if the teachers had also used the criteria to provide students with written evaluations of their research reports. Unfortunately,
because the teachers taught as many as 150 students each (five classes of approximately 30 students), they did not have time to judge and characterize their students' work on each of these 10 criteria. Instead, they gave students much more limited feedback, usually just a letter grade (or even just a check mark) with no written comments. (The teachers thus did not evaluate the projects using the scoring criteria until the following summer as part of the scoring study described earlier.) To help remedy this problem, we have subsequently developed a computer-based scoring tool, which teachers can use to generate a written evaluation as the students give oral presentations of their research work to the class.

In this tool, shown in Figure 25, each of the criteria for characterizing scientific research are presented on the screen. The teachers can click on the name of any criterion and see its definition. If they click on the icon for a given criterion, a 5-point scale appears with a written characterization for each of the levels in the 5-point scale. The teachers can then select the point on the scale that they feel characterizes the students research. If they do this for each of the criteria, the software generates a written report that provides the student with a score and corresponding level descriptor for each of the criteria. The teachers can edit these descriptors if the ones given do not characterize the student's research appropriately. We found that the process of working with teachers to decide what each level descriptor should say was very worthwhile in terms of helping teachers to think about the characteristics of good research. If this reflective-assessment tool had been available for the instructional study presented in this article, the effects of the Reflective-Assessment Process might have been even greater. Certainly providing both students and teachers with this tool, and having students participate in

<table>
<thead>
<tr>
<th>Student's Name</th>
<th>Barbara</th>
<th>Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module</td>
<td></td>
<td>Period</td>
</tr>
<tr>
<td>Project #</td>
<td></td>
<td>Date</td>
</tr>
<tr>
<td>Understanding the Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="3" alt="Lightbulb" /> +</td>
<td>Shows understanding of basic concepts but with little depth.</td>
<td></td>
</tr>
<tr>
<td>Understanding Inquiry</td>
<td><img src="3" alt="Question Mark" /> +</td>
<td>Inquiry cycle is followed, but with little depth in some phases.</td>
</tr>
<tr>
<td>Experimental Design</td>
<td><img src="4" alt="Exclamation" /> +</td>
<td>Well designed expts. for both computer and real world.</td>
</tr>
<tr>
<td>Communicating Well</td>
<td><img src="3" alt="Envelope" /> +</td>
<td>Explains experiment clearly, but with no depth to discussion.</td>
</tr>
<tr>
<td>Overall Score</td>
<td>3</td>
<td>Barbara, you did a good job on your first big project!</td>
</tr>
</tbody>
</table>

FIGURE 25 The computer-based scoring tool that we developed to enable teachers to more easily use the criteria for judging research when assessing students' research projects.
conversations about what each level descriptor should say, would be an interesting pedagogical experiment.

*Problems Raised by the Students Themselves*

The students had two major complaints about their experiences with the ThinkerTools Inquiry Curriculum. The first can be illustrated by the following quote:

Too much self-assessment! You asked us to assess our work at the end of each and every step in the Inquiry Cycle, as well as at the end of each module where you also asked us to assess how well we assessed ourselves. Don’t you think that’s a bit much?

Such complaints may indicate that students do not appreciate the value of the Reflective-Assessment Process, or they may simply reveal a genuine overdose. To investigate this issue, we modified the curriculum to reduce the amount of self-assessment. At the end of each step in the Inquiry Cycle, students now evaluate their work on only the single, most-relevant scoring criterion. At the end of each module, they give their work an overall score and justify it. They are no longer asked to reflect on each of the scoring criteria or on the self-assessment process itself. From this study, we will determine if the same improvement in inquiry skills occurs and if students value rather than complain about the self-assessment—reports from the teachers indicate that they do value it.

The second major complaint students had about the curriculum can be illustrated by the following quote:

Too much repetition. We have to write down our theory when we make our predictions, and we have to write it down again after doing our experiments. We also have to write a plan for our experiment, and then we have to write a research report that describes it again. Too much writing of the same thing over and over again.

A word processor would clearly improve the situation, particularly the copy command, and should reduce such complaints. However, these complaints may indicate that some students do not appreciate the importance of revising their theories and research plans based on data and reflection. This may suggest a need to introduce another metacognitive level into such curricula, one that is aimed at helping students develop knowledge about the importance of metacognition. We are pursuing this idea in some of our present work by having students engage in cognitive research following their physics research. In this follow-on curriculum, they create and test theories about the process of theory development itself.
Summary of Challenges

In summary, one of the main challenges in creating this curriculum has been to create inquiry and reflection activities that are motivating and meaningful to students, develop their inquiry and modeling skills, and are easy for teachers to orchestrate. In our revisions of the curriculum, we need to be concerned with students' acceptance of the importance of the research notebooks and reflection activities and to encourage more critical evaluation of their research results. We need to ensure that they view themselves as a community of researchers who are trying to construct a common, shared theory of force and motion.

Dissemination and Teacher Professional Development

A challenge for the dissemination of curricula, like the ThinkerTools Inquiry Curriculum, is that few teachers, particularly at the elementary and middle school levels, know inquiry or physics, and so they may have difficulty teaching it. However, the view that teachers need to be experts in physics in order to teach it is derived primarily from didactic approaches to instruction. If, instead, one takes a constructivist, inquiry-oriented approach, then such limitations on who can teach science may no longer apply. Our initial hypothesis was that teachers with no prior inquiry background and no prior subject matter background should be able to teach science successfully if they are provided with a constructivist curriculum that includes the appropriate supporting materials: These include materials that scaffold the learning of scientific inquiry as well as computer simulations that facilitate the discovery of the laws of physics (by providing simplified, readily controllable microworlds from which it is relatively easy to induce accurate laws). Our view was that such curricular and computer-based materials should enable teachers to successfully implement an inquiry-oriented approach and, in so doing, to learn about physics and scientific inquiry and how to teach it. Because the study reported in this article included only three teachers, it provided only preliminary, albeit positive, evidence with regard to this hypothesis.

The limitations of this initial hypothesis were revealed, however, in more recent and extensive trials of the ThinkerTools Inquiry Curriculum in which it was sent "mail order" to eight teachers (all at a great distance from us). We had no contact with any of these teachers and simply asked them to fill out questionnaires asking them to outline their goals for science teaching and to describe what they did and did not like about ThinkerTools. They were also asked to give their students our Inquiry Test and our Applied Physics Test before and after they completed the curriculum. These data revealed some interesting findings. Four of the teachers said that their primary goal was to teach scientific inquiry and described the ThinkerTools materials as an exciting way to teach inquiry. For the students of these
teachers, we found the same significant gains in physics and inquiry knowledge as
in the study reported in this article. The other four teachers said, in contrast, that
their focus was on teaching physics, that is, teaching the subject matter, and
described ThinkerTools as an exciting way to teach physics. For the students of
these teachers, we found significant gains on our Applied Physics Test but not on
our Inquiry Test. When we later asked the teachers whether they taught all of the
curriculum, we found that the "physics-focused" teachers dropped many of the
inquiry components.23

Furthermore, when we analyzed videotapes of six local teachers teaching
ThinkerTools, we found that the ways in which they implemented the curriculum
often did not match what we envisioned with regard to good inquiry teaching. For
example, in the Prediction Phase of the Inquiry Cycle, when the teachers were
trying to get students to articulate a set of competing hypotheses, they often did
not ask students to justify their predictions with some form of causal explanation.
Also, when students were having difficulty in designing and conducting their
experiments, the teachers sometimes gave a procedural solution without getting
the students to analyze the design problem for themselves. Furthermore, in the
Model Phase of the Inquiry Cycle, the goal was to try to reach a consensus about
the law(s) that best characterize what the students had discovered from their
experiments. Here, some teachers used aspects of suggested activities, such as
asking students to vote on the best law, but did not ask their students to cite theory
and evidence to support their choice or to engage in debates using evidence from
their experiments.

Based on these findings, we determined that it is not enough to simply provide
teachers with Teacher’s Guides that attempt to outline goals, suggest activities, and
describe, in a semiprocedural fashion, how the lessons might proceed. We now
argue that, in addition, teachers need to develop a conceptual framework for
characterizing good inquiry teaching and for reflecting on their teaching prac-
tices—in the same way that students need to develop criteria for characterizing
good scientific research and for reflecting on their inquiry processes.

To achieve the goal of enabling teachers to characterize and reflect on inquiry
teaching, we are now using a framework developed for the National Board for
Professional Teaching Standards (Frederiksen et al., 1997). This framework, which
attempts to characterize expert teaching, includes five major criteria: worthwhile
engagement, adept classroom management, effective pedagogy, good classroom

23The present curriculum takes approximately 10 weeks to complete. Most of the middle school
science teachers we have worked with want curricular units that take only 3 to 6 weeks. Our present
curriculum thus invites teachers who are not aficionados of extended inquiry to make cuts. Also, asking
a teacher to try a new approach for as much as 10 weeks may be a little unrealistic. We are thus presently
working on various shorter versions of the curriculum that, nonetheless, incorporate the same focus on
inquiry.
climate, and thinking about the subject matter, to which we added engaging in inquiry. In this characterization of expert teaching, each of these criteria for good teaching is unpacked into a set of “aspects.” For example, Figure 26 illustrates the criterion of “good classroom climate,” which is defined as “the social environment of the class empowers learning.” Under this general criterion, there are five different aspects: engagement, encouragement, rapport, respect, and sensitivity to diversity. Each of these aspects is defined in terms of specific characteristics of classroom practice, such as “humor is used effectively” or “there is a strong connection between students and teacher.” Furthermore, each of these specific characteristics of classroom practice is indexed to a set of video clips, called “video snippets,” which illustrate it. This framework characterizes good inquiry teaching and provides teachers with video exemplars of teaching practice.

Such materials can be used to enable teachers to learn about inquiry teaching and its value as well as to reflect on their own and each others’ teaching practices. For example, recently we tried the following approach with a class of 10 student teachers. The student teachers first learned to use the framework outlined in the preceding paragraph by scoring some of our ThinkerTools videotapes. Then, they used the framework to facilitate discussions of videotapes of their own teaching.

---

**FIGURE 26** An example of the hierarchical definitions created for each of the five criterion, such as classroom climate, which are used to characterize expert teaching (from Frederiksen, Sipusic, Gamoran, & Wolfe, 1997).
In this way, they participated in what we call “video clubs,” which enable them to reflect on their own teaching practices and to hopefully develop better approaches for inquiry teaching. (Video clubs incorporate social structures designed to help teachers reflectively assess and talk about their teaching practices.) The results so far have been very encouraging (Frederiksen & White, 1997a).

We thus argue that the same emphases on inquiry, scaffolding, metacognition, and reflection that we have found are important and effective for students are important for teachers as well. This can partly be achieved by providing teachers with teacher’s guides that scaffold each step in the Inquiry Cycle—outlining the pedagogical goals and describing the instructional activities and how they can be introduced. The teacher’s guides, however, are not enough. In addition, teachers need to be introduced to a reflective-assessment process in which they learn about the characteristics of good inquiry teaching and learn to reflect on their teaching practices. We are presently conducting research with in-service as well as preservice teachers to further investigate these hypotheses about how to enable teachers to adopt inquiry-oriented approaches to their teaching.

SUMMARY AND CONCLUSIONS

In this project, we created the ThinkerTools Inquiry Curriculum, which incorporates a constructivist, inquiry-oriented approach to science education in which the development of metacognitive knowledge and skills plays a central role. We then analyzed the effects that this curriculum had on students who varied in their degree of educational advantage, as measured by their grade levels and standardized achievement test scores. The students were typical students taught by regular classroom teachers in urban, public, middle schools. We compared the performance of these middle schools students with that of high school physics students. We also carried out a controlled study comparing ThinkerTools classes who engaged in a Reflective-Assessment Process with matched control classes who did not. In the reflective-assessment classes, the students were continually engaged in monitoring and evaluating their own and each other’s research. All classes participated in the same ThinkerTools inquiry-based science curriculum in which they designed and carried out experiments using hands-on materials and computer simulations and then developed laws, models, and theories to account for their findings.

Our results show that software modeling tools—which enable students to work with and construct conceptual models that use diagrammatic representations and employ causal reasoning in which they step through time to analyze events—make the difficult subject of physics understandable and interesting to a wide range of students. Furthermore, the focus on creating models enables students to learn not only about physics but also about the properties of scientific models and the inquiry processes needed to create them. We found that, regardless of their lower grade
levels (7–9) and their lower pretest scores, students who had participated in ThinkerTools outperformed high school physics students (Grades 11–12) on qualitative problems in which they were asked to apply the basic principles of Newtonian mechanics to real-world situations. In general, this inquiry-oriented, model-based, constructivist approach to science education appears to make science interesting and accessible to a wider range of students than is possible with traditional approaches.

The results of our controlled comparison revealed that the students' learning of inquiry was greatly facilitated by introducing the Reflective-Assessment Process. We found that incorporating the Reflective-Assessment Process had the effect of increasing the quality of students' research projects and inquiry test performance over that of students in the control classes, even though the methods of scientific inquiry were carefully developed for both groups.

An important finding was that this beneficial effect of metacognitive reflection was particularly strong for the lower achieving students: The Reflective-Assessment Process enabled the lower achieving students to gain more on inquiry tests. It also enabled them to perform close to the higher achieving students on their research projects. The introduction of reflective assessment, although helpful to all, was particularly helpful in closing the performance gap between the lower and higher achieving students. In fact, the Reflective-Assessment Process enabled lower achieving students to perform as well as higher achieving students on their research projects if they did their research in collaboration with a higher achieving student. In the control classes, in contrast, the lower achieving students did not do as well as high-achieving students, regardless of whether they collaborated with a higher achieving student. Also, the beneficial effect of the Reflective-Assessment Process for the high-achieving students was the same for those who collaborated with a low-achieving student as for those who collaborated with a high-achieving student.

These results suggest that, from an equity standpoint, curricular and assessment approaches can be created that are not merely equal in their value for, but actually enhance, the learning of less advantaged students, without impeding the high-achieving students. We think that these findings have strong implications for what such inquiry-oriented, metacognitively focused curricula can accomplish, particularly in an urban school setting in which there are many disadvantaged students.

Transferable Inquiry and Metacognitive Expertise

What conceptual and metacognitive knowledge and skills do students derive from their work in the ThinkerTools Inquiry Curriculum? The teachers report and our research indicates that they acquire an understanding of the Inquiry Cycle (see Figure 1) as well as the knowledge needed to carry out each step in this cycle. They
also acquire knowledge of the forms that scientific laws, models, and theories can take and of how the development of scientific theories is related to empirical evidence. In addition, they acquire the metacognitive skill of monitoring and reflecting on their inquiry (see the end of Appendix B for an example). Because all of science can be viewed as a process of constructing models and theories, both the Inquiry Cycle and the Reflective-Assessment Process can be applied to learning and doing all areas of science, not just physics. Thus understanding and engaging in the Inquiry Cycle and Reflective-Assessment Process should benefit students in their future science courses.

We see evidence of these benefits and transfer in the subsequent work of ThinkerTools students. For example, one cohort of seventh-grade students used the Inquiry Cycle to guide their research for their Science Fair Projects. These projects were done after the ThinkerTools Inquiry Curriculum and included a wide range of topics. Only 3 of the 8 seventh-grade science classes in the school had done ThinkerTools, and yet, ThinkerTools students won 10 out of the 13 awards (the students' projects were judged by scientists from the local community). When these students reached the eighth grade, their teacher asked them to do a research project by following the Inquiry Cycle. They were free to choose topics other than physics. For instance, one group of students wanted to understand how listening to music affects one's performance on schoolwork. This group did an experiment in which they had their classmates listen to different kinds of music while taking arithmetic tests. The students wrote research reports that described how they planned and carried out their research, and they evaluated their own and each other's research using the criteria shown in Figure 8. Their teacher reports that their performance on these projects was equal to or better than the performance on their ThinkerTools physics projects. Furthermore, at the end of the curriculum, some students were interviewed and asked if the Inquiry Cycle and Reflective-Assessment Processes could be used to help them learn other subjects. Many of their answers involved highly creative explanations of how they could be applied to domains such as history, mathematics, and English, as well as to other areas of science. With regard to the teachers, all of them attest to the benefits of both the Inquiry Cycle and the Reflective-Assessment Process and have chosen to incorporate them into the other science courses that they teach.

In order to make the valuable skills of inquiry, modeling, and reflection apply to other experimental sciences, such as biology, as well as to the learning of nonscience subjects, various approaches could be pursued. For instance, students could be introduced to a generalized version of the Inquiry Cycle (such as Question, Hypothesize, Investigate, Analyze, Model, and Evaluate — this represents a minor transformation of the more experimentally oriented Inquiry Cycle that students internalize during the ThinkerTools Inquiry Curriculum). This generalization could give students a metacognitive view of learning and inquiry that can be applied to any topic in which building predictive-explanatory models can become the focus.
In addition, the students could discuss how the Reflective-Assessment Process that uses the criteria shown in Figure 8 (such as Making Connections, Reasoning Carefully, and Communicating Well) can readily be generalized to learning other science topics as well as to learning in general. Having such explicit discussions of transfer in conjunction with explicitly using versions of the Inquiry Cycle and Reflective-Assessment Process in their science and other curricula should enable students and teachers to appreciate and benefit from the power of metacognition. Investigating how such generalization and transfer can be achieved will be a major focus of our future research.

ACKNOWLEDGMENTS

This research was supported by James S. McDonnell Foundation Award 9135, National Science Foundation Award DPE–8400280 with the Directorate for Science and Engineering Education, and the Educational Testing Service. We appreciate the interest, support, and encouragement of John Bruer of the McDonnell Foundation and Henry Braun of the Educational Testing Service, and we are grateful to the Oakland and Berkeley Public Schools for allowing us to conduct this research. We also thank Isabel Beck, Allan Collins, Bob Glaser, Leona Schauble, Karen Shiengold, and several anonymous reviewers for their helpful comments on an earlier version of this article.

The ThinkerTools Inquiry Project is a collaborative endeavor between researchers at the University of California at Berkeley and the Educational Testing Service and middle school teachers in the Berkeley and Oakland public schools. The team includes principal investigators Barbara White and John Frederiksen; teachers Vana James, Din Seaver, and Linda Taylor White; research assistants Sue Allen, Joshua Gutwill, Christine Schwarz, Noel Enyedy, Lisa Pino, and Todd Shimoda; programmers Laura Werner and Christopher Schneider; and secretary Cynthia Sultan. We are grateful for the contributions and hard work of each member of the team. We also acknowledge the foundational research of the prior Thinker-Tools project (principle investigators Paul Horwitz and Barbara White), in which earlier versions of these curricular materials and software were created and tried out in sixth-grade classes (see Horwitz, 1989; B. White, 1993b; B. White & Horwitz, 1988).

REFERENCES


APPENDIX A
Examples of Scaffolded Inquiry Activities

(from Module 1)

Evaluating Laws

It is useful to summarize what you discovered from working with the computer model with a law that describes what happens when you apply impulses to the dot.

Basic Principles

For each of the following laws, decide if you think it correctly describes what happens. After you have evaluated all of the laws, pick the one that you think best states the basic principle of the computer model.

1. Whenever you apply an impulse to the dot, it changes speed.

Is this law correct or incorrect? (Circle your choice below)

- correct
- incorrect

Explain your reasoning. ____________________________________________

__________________________________________

__________________________________________

2. The dot changes speed whenever it wants to.

Is this law correct or incorrect? (Circle your choice below)

- correct
- incorrect

Explain your reasoning. ____________________________________________

__________________________________________

__________________________________________
3. Whenever you give the dot an impulse, it speeds up.

Is this law correct or incorrect? (Circle your choice below)

**Correct**  
**Incorrect**

Explain your reasoning. ____________________________________________  
_________________________________________________________________  
_________________________________________________________________

4. The dot's speed stays the same except when you apply an impulse.

Is this law correct or incorrect? (Circle your choice below)

**Correct**  
**Incorrect**

Explain your reasoning. ____________________________________________  
_________________________________________________________________  
_________________________________________________________________

Which law do you think best states the basic principle of the computer model? (Circle your choice)

1  2  3  4

Explain your reasoning. ____________________________________________  
_________________________________________________________________  
_________________________________________________________________  
_________________________________________________________________
More Precise Laws

A useful law is a simple, easy-to-remember rule that enables you to predict what will happen in many different situations when you are working with the computer model. For instance, a useful law would enable you to predict precisely what will happen if you give the dot a sequence of impulses such as \((\Rightarrow, \Rightarrow, \Leftarrow)\) or such as \((\Leftarrow, \Rightarrow, \Rightarrow, \Leftarrow)\).

For each of the following laws, describe how useful it is. After you have evaluated all of the laws, pick the one that is the most useful.

5. If the dot is moving to the right and you apply an impulse to the right, the dot will speed up.

Is this law useful for predicting what will happen? (Circle your choice)

not useful somewhat useful very useful

Explain your reasoning.

6. If you keep applying impulses in the direction that the dot is moving, it keeps speeding up. If you keep applying impulses in the direction opposite to which the dot is moving, it slows down, stops, and goes the other way.

Is this law useful for predicting what will happen? (Circle your choice)

not useful somewhat useful very useful

Explain your reasoning.
7. If the dot is moving to the left and you apply an impulse to the right, the dot will slow down.

Is this law useful for predicting what will happen? (Circle your choice)

not useful  somewhat useful  very useful

Explain your reasoning. ____________________________________________

_________________________________________________________________

_________________________________________________________________

8. If you apply an impulse in the same direction that the dot is moving, it adds 1 to its speed. If you apply an impulse in the direction opposite to which the dot is moving, it subtracts 1 from its speed.

Is this law useful for predicting what will happen? (Circle your choice)

not useful  somewhat useful  very useful

Explain your reasoning. ____________________________________________

_________________________________________________________________

_________________________________________________________________

Which law do you think is the most useful? (Circle your choice)

5  6  7  8

Explain your reasoning. ____________________________________________

_________________________________________________________________

_________________________________________________________________
Two students were interested in finding out what friction does to the motion of balls. They followed the inquiry cycle and designed and carried out an experiment. Then one of them wrote the research report shown on the next two pages. Unfortunately, there are a lot of problems with what they did. How many things can you find wrong with their research?

1. Describe and explain each problem that you find with their research. (See if you can find at least 5.)


2. Now that you have described the problems with their research, describe what you think is good about it.


3. Design your own experiment to investigate the effects of friction. Try to avoid the problems you found with the experiment you just criticized.
Research Report on Friction and Motion

Question:

How does friction affect motion?

Predict:

My hypothesis was that a ball will go faster on a smooth surface than on a rough surface because I know that is what happens.

My partner thought that the surface will not make any difference because you can control how fast a ball goes by how hard you hit it.

Experiment:

For our experiment we used a tennis ball, a billiard ball, a bonker, and a meter stick. We put the tennis ball on the floor and bonked it and measured how far it went. Then we put the billiard ball on the carpet and bonked it and measured how far it went.
Our Data Table

<table>
<thead>
<tr>
<th>Surface</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>tennis ball on floor</td>
<td>3.55 meters</td>
</tr>
<tr>
<td>billiard ball on carpet</td>
<td>3.52 meters</td>
</tr>
</tbody>
</table>

Model:

Our data show that the tennis ball went further than the billiard ball.

From this we concluded that my hypothesis was right.

The law we discovered is that objects go further on smooth surfaces than on rough surfaces.

We think this happens because rough surfaces are bumpy.

Apply:

What we learned could be useful. For example, if you wanted to teach someone to play hockey, you could have them play on a rough surface like grass instead of on a smooth surface like ice, because the ball will go slower and it will be easier.

The main problem with our experiment is that we should have tried bonking more kinds of balls on more kinds of surfaces. If we had time to do more research, we would try bonking a bowling ball on a hill and a beach ball on a sandy beach.
APPENDIX B
Instructions for the Mass Project and an Example of a
Student’s Project Report and Self-Assessment

Instructions for Module 3

Mass and the Effects of an Impulse:
What is the Relationship?

In this module, you and your partner will do a research project together. The goal is to investigate how the mass of an object (how heavy it is) affects what an impulse does to its motion. In other words, you want to find out how objects with different masses (different weights) respond to standard impulses. You will design both computer and real-world experiments to study this. Your investigation should follow the inquiry cycle as closely as possible.

Computer experiments. You can use the ThinkerTools software to create computer models and experiments. A guide to how to do this can be found at the end of this module. In the guide, you will find out how to construct a ThinkerTools model with more than one dot, and how to add dotprints, datacrosses, timers, and a data table to help you see what happens to the dot when you apply an impulse to it. You will also find out how to show and change the mass of any dot.

Real-world experiments. You can use any of the following lab equipment for your real-world experiments:

- bonkers
- meter sticks
- stopwatches
- scales
- balls
- rolling carts
- plastic pucks
- weights
Steps for Doing Your Project:

1. **Write a Research Plan.** In this plan, clearly state your research question and hypotheses. Then describe the computer and real-world experiments that you will do to investigate this question. Check your plan to make sure that it meets the following requirements:
   (a) your research can be done in the time allowed;
   (b) what you are going to do is clearly thought out; and
   (c) the experiments will allow you to answer the research question.

2. **Keep a Research Journal.** Carry out your research. Record in your journal what you do each day, any results you have, and any conclusions you make, as well as your reasons for what you are doing and thinking. If you find that you need to revise your research plan as you go along, record the changes you have made and the reasons for them.

3. **Write a Project Report.** You and your partner should each write your own reports. Your Project Report should show how you followed the inquiry cycle to investigate your research question. An outline and checklist for your report are presented on Page 4. A title page is provided on page 5. Use your Research Plan and your Research Journal to help you write your Project Report.

4. **Give an Oral Presentation to the Class.** Your goal is to explain to the class what you did at each step in the inquiry cycle. Your presentation may include a demonstration of computer and real-world experiments. It should also include a poster to show your results. Be prepared to answer questions from the class.

5. **Evaluate Your Project.** Your project will be evaluated by your teacher, yourself (you will fill out the Scorecard for Judging Project Work on pages 8-11), and by other students in the class (you will get to evaluate their projects as well). The projects will be evaluated based on the Guidelines for Judging Project Work shown on pages 6-7.
An Outline and Checklist for Your Project Report

☐ Question:
  • Clearly state the research question.

☐ Predict:
  • What hypotheses did you have about possible answers to the question?
    • Explain the reasoning behind each of your hypotheses.

☐ Experiment:
  • Describe your computer experiment(s).
    • Draw a sketch of your computer model.
    • Describe how you used it to carry out your experiment(s).
  • Show your data in tables, graphs, or some other representation.
  • Describe your real-world experiment(s).
    • Draw a sketch of how you set up the lab equipment.
    • Describe how you used the equipment to carry out your experiment(s).
  • Show your data in tables, graphs, or some other representation.

☐ Model:
  • Describe how you analyzed your data and show your work.
  • Summarize your conclusions.
    • Which of your hypotheses does your data support?
    • State any laws that you discovered.
  • What is your theory about why this happens?

☐ Apply:
  • Show how what you learned could be useful.
    • Give some examples.
  • What are the limitations of your investigation?
    • What remains to be learned about the relationship between the mass of an object and how forces affect its motion?
    • What further investigations would you do if you had more time?
During the past few weeks, my partner and I have been creating and doing experiments and making observations about mass and motion. We had a specific question that we wanted to answer -- how does the mass of a ball affect its speed?

I made some predictions about what would happen in our experiments. I thought that if we had two balls of different masses, the ball with the larger mass would travel faster, because it has more weight to roll forward with, which would help push it.

We did two types of experiments to help us answer our research question -- computer and real world. For the computer experiment, we had a ball with a mass of 4 and a ball with a mass of 1. In the real world they are pretty much equal to a billiard ball and a racquetball. We gave each of the balls 5 impulses, and let them go. Each of the balls left dotprints, that showed how far they went for each time step. The ball with the mass of 4 went at a rate of 1.25 cm per time step. The ball with the mass of 1 went at a rate of 5 cm per time step, which was much faster.

For the real world experiment, we took a billiard ball (with a mass of 166 gms) and a racquetball (with a mass of 40 gms). We bonked them once with a rubber mallet on a linoleum floor, and timed how long it took them to go 100 cm. We repeated each experiment 3 times and then averaged out the results, so our data could be more accurate. The results of the two balls were similar. The racquetball's average velocity was 200 cm per second, and the billiard ball's was 185.1 cm per second. That is not a very significant difference, because the billiard ball is about 4.25 times more massive than the racquetball.
We analyzed our data carefully. We compared the velocities, etc. of the lighter and heavier balls. For the computer experiment, we saw that the distance per time step increased by 4 (from 1.25 cm to 5 cm) when the mass of the ball decreased by 4 (from 4 to 1). This shows a direct relationship between mass and speed. It was very hard to analyze the data from our real world experiment. One reason is that it varies a lot for each trial that we did, so it is hard to know if the conclusions we make will be accurate. We did discover that the racquetball, which was lighter, traveled faster than the billiard ball, which was heavier.

Our data doesn’t support my hypothesis about mass and speed. I thought that the heavier ball would travel faster, but the lighter one always did. I did make some conclusions. From the real world experiment I concluded that the surface of a ball plays a role in how fast it travels. This is one of the reasons that the two balls had similar velocities in our real world experiment. (The other reason was being inaccurate). The racquetball’s surface is rubbery and made to respond to a bonk and the billiard ball’s surface is slippery and often makes it roll to one side. This made the balls travel under different circumstances, which had an effect on our results.

From the computer experiment I concluded that a ball with a smaller mass goes as many times faster than a ball with a larger mass as it is lighter than it. This happens because there is a direct relationship between mass and speed. For example, if you increase the mass of a ball then the speed it travels at will decrease.

I concluded in general, of course, that if you have two balls with different masses that the lighter one will go faster when bonked, pushed, etc. This is because the ball doesn’t have as much mass holding it down.

The conclusions from our experiments could be useful in real world experiences. If you were playing baseball and you got to choose what ball to use, you would probably choose one with a rubbery surface that can be gripped, over a slippery, plastic ball. You know that the type of surface that a ball has effects how it responds to a hit. If you were trying to play
catch with someone you would want to use a tennis ball rather than a billiard ball, because you know that balls with smaller masses travel faster and farther.

The investigations that we did do have limitations. In the real world experiments the bonks that we gave the balls could have been different sizes, depending on who bonked the ball. This would affect our results and our conclusions. The experiment didn’t show us how fast balls of different masses and similar surfaces travel in the real world. That is something we still can learn about. If there was more time, I would take two balls of different masses with the same kind of surface and figure out their velocities after going 100 cm.

Overall, our experiments were worthwhile. They proved an important point about how mass affects the velocity of a ball. I liked being able to come up with my own experiments and carrying them out.
COMPUTER
EXPERIMENTS

<table>
<thead>
<tr>
<th>MASS</th>
<th>DIST. PER TIME STEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.25 cm.</td>
</tr>
<tr>
<td>1</td>
<td>5 cm.</td>
</tr>
</tbody>
</table>

\[ \text{mass: 1} \]

\[ \text{mass: 4} \]
### RACQUETBALL

<table>
<thead>
<tr>
<th>Mass: 40 g.</th>
<th>DIST.</th>
<th>TIME</th>
<th>VELO.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TRIAL 1</strong></td>
<td>100 cm.</td>
<td>.56 sec</td>
<td>178.5 cm per sec</td>
</tr>
<tr>
<td><strong>TRIAL 2</strong></td>
<td>100 cm.</td>
<td>.43 sec</td>
<td>232.5</td>
</tr>
<tr>
<td><strong>TRIAL 3</strong></td>
<td>100 cm.</td>
<td>.53 sec</td>
<td>189.0</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td>100 cm.</td>
<td>.51 sec</td>
<td>200.0</td>
</tr>
</tbody>
</table>

### BILLIARD BALL

<table>
<thead>
<tr>
<th>Mass: 166 g.</th>
<th>DIST.</th>
<th>TIME</th>
<th>VELO.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TRIAL 1</strong></td>
<td>100 cm.</td>
<td>.60</td>
<td>166.6 cm per sec</td>
</tr>
<tr>
<td><strong>TRIAL 2</strong></td>
<td>100 cm.</td>
<td>.50</td>
<td>200</td>
</tr>
<tr>
<td><strong>TRIAL 3</strong></td>
<td>100 cm.</td>
<td>.53</td>
<td>188.7</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td>100 cm.</td>
<td>.54</td>
<td>185.1</td>
</tr>
</tbody>
</table>
We bonked balls of two different masses and timed how long it took for them to go 100 cm. and then figured out their velocities.
An example of a self-assessment produced by the 7th grade student who did the preceding Mass Project

**UNDERSTANDING**

Understanding the Science

<table>
<thead>
<tr>
<th>NA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4v</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>adequate</td>
<td>not adequate</td>
<td>adequate</td>
<td></td>
<td></td>
<td>exceptional</td>
</tr>
</tbody>
</table>

Justify your score based on your work. I have a basically clear understanding of how mass affects the motion of a ball in general, but I don't have a completely clear sense of what would happen if friction, etc. was taken into account.

Understanding the Processes of Inquiry

<table>
<thead>
<tr>
<th>NA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4v</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>adequate</td>
<td>not adequate</td>
<td>adequate</td>
<td></td>
<td></td>
<td>exceptional</td>
</tr>
</tbody>
</table>

Justify your score based on your work. I used the inquiry cycle a lot in my write up, but not as much while I was carrying out my experiments.

Making Connections

<table>
<thead>
<tr>
<th>NA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4v</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>adequate</td>
<td>not adequate</td>
<td>adequate</td>
<td></td>
<td></td>
<td>exceptional</td>
</tr>
</tbody>
</table>

Justify your score based on your work. I made some references to the real world, but I haven't fully made the connection to everyday life.
PERFORMANCE: DOING SCIENCE

Being Inventive

<table>
<thead>
<tr>
<th>NA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4\dagger</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not adequate</td>
<td>adequate</td>
<td>exceptional</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Justify your score based on your work. What I did was original, but many other people were original and did the same (or similar) experiment as us.

Being Systematic

<table>
<thead>
<tr>
<th>NA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4\dagger</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not adequate</td>
<td>adequate</td>
<td>exceptional</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Justify your score based on your work. On the whole I was organized, but if I had been more precise my results would have been a little more accurate.

Using the Tools of Science

<table>
<thead>
<tr>
<th>NA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5\dagger</th>
</tr>
</thead>
<tbody>
<tr>
<td>not adequate</td>
<td>adequate</td>
<td>exceptional</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Justify your score based on your work. I used many of the tools I had to choose from. I used them in the correct way to get results.

Reasoning Carefully

<table>
<thead>
<tr>
<th>A \rightarrow B</th>
<th>B \rightarrow C</th>
<th>A \rightarrow C</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>not adequate</td>
<td>adequate</td>
<td>exceptional</td>
</tr>
</tbody>
</table>

Justify your score based on your work. I took into account the surfaces of the balls in my results, but I didn't always reason carefully. I had to ask for help, but I did compute out our results mathematically.
Writing and Communicating Well

Justify your score based on your work. I understand the science, but in my writing and comments I might have been unclear to others.

Teamwork

Justify your score based on your work. We got along fairly well and had a good project as a result. However, we had a few arguments.

Self-assessment

How well do you think you evaluated your work using this scorecard?

I think I judged myself fairly - not too high or too low. I didn't always refer back to specific parts of my work to justify my score.