Development of Scientific Reasoning in College Biology: Do Two Levels of General Hypothesis-Testing Skills Exist?

Anton E. Lawson,[1] Brian Clark,[1] Erin Cramer-Meldrum,[1]
Kathleen A. Falconer,[1] Jeffrey M. Sequist,[1] Yong-Ju Kwon[2]

[1]*Department of Biology, Arizona State University, Tempe, Arizona 85287-1501*

[2]*Center for Innovative Teaching and Learning, Pohang University of Science and Technology, Pohang, Kyungbuk 790-784, Korea*

Received 22 May 1998; accepted 14 September 1999

Abstract: The primary purpose of the present study was to test the hypothesis that two general developmentally based levels of hypothesis-testing skills exist. The first hypothesized level presumably involves skills associated with testing hypotheses about observable causal agents; the second presumably involves skills associated with testing hypotheses involving unobservable entities. To test this hypothesis, a hypothesis-testing skills test was developed and administered to a large sample of college students both at the start and at the end of a biology course in which several hypotheses at each level were generated and tested. The predicted positive relationship between level of hypothesis-testing skill and performance on a transfer problem involving the test of a hypothesis involving unobservable entities was found. The predicted positive relationship between level of hypothesis-testing skill and course performance was also found. Both theoretical and practical implications of the findings are discussed. © 2000 John Wiley & Sons, Inc. J Res Sci Teach 37: 81–101, 2000

Following the review of several years of research into problem-solving performance, Perkins and Salomon (1989) concluded that, although expert performance manifests itself in contextualized ways, general cognitive skills (i.e., "habits of mind") exist. These general cognitive skills reveal themselves primarily as strategies of looking for counterexamples to test causal knowledge claims. Although Perkins and Salomon discussed such strategies as thinking tools of the philosopher, scientists recognize them as components of a scientific method that has as its core the generation and test of alternative hypotheses (cf. Baker & Allen, 1977; Burmester, 1952; Carey, 1998; Chamberlain, 1965; Lawson, 1995; Lewis, 1988; Moore, 1993; Platt, 1964). Essentially, this method embodies a set of generally applicable questions that must be raised and satisfactorily answered before drawing a firm conclusion about the relative truth or falsity of any particular causal claim. The set of questions reads more or less like this: What is the central causal question raised in this particular context? In addition to the proposed cause, what al-

Correspondence to: A. E. Lawson
Contract grant sponsor: National Science Foundation
Contract grant number: DUE 9453610

ternative causes (i.e., hypotheses/theories) are possible? How can each possibility be tested? What specific expectations (i.e., predictions) follow from each possibility and its proposed test? How does the evidence (either circumstantial, correlational, or experimental), once gathered, match the expectations? What conclusions can be drawn based on the obtained degree of match?

Of course, the development of reasoning patterns associated with these questions has been the subject of a long line of research within developmental psychology and within science education (for reviews, see Lawson, 1985, 1992a; for more recent research within science education, see, for example, Cavallo, 1996; Germann, 1994; Germann & Aram, 1996; Hurst & Milkent, 1996; Johnson & Lawson, 1998; Keys, 1994; Kuhn, 1989; Lawson, 1992b; Lawson & Thompson, 1988; Lawson & Worsnop, 1992; Noh & Scharmann, 1997; Shayer & Adey, 1993; Westbrook & Rogers, 1994; Wong, 1993; Zohar, Weinberger, & Tamir, 1994). The general conclusion of such research is that reasoning patterns (the exact nature of which is yet to be determined) do develop across adolescence, at least in some students, and play an important role in the ability to do science and to construct science concepts. Research has also documented that improvements in reasoning as a consequence of instruction, although difficult to obtain, are possible and of general use. Dramatic evidence of this was obtained by Shayer and Adey (1993), who found that 3 years after the end of a 2-year science program designed to promote formal operational thinking, positive effects were seen on the British National examinations not only in science, but also in mathematics and English.

Consequently, one of the goals of our department's introductory nonmajors' biology course, The Living World, is to help students develop generalizable hypothesis-testing skills by encouraging them to raise and answer the previously listed set of questions during a series of lab and field activities. In addition, the course lectures present several episodes that explicate how the questions have been asked and answered by biologists while conducting past research. In other words, the course attempts to teach general cognitive skills essentially in the way described as the "high road" by Perkins and Salomon, which is say that the course encourages the "deliberate and mindful abstraction" of the question-asking and question-answering skills from a variety of domain-specific contexts.

During a recent semester, the quizzes listed in Table 1 were administered as part of an effort to assess the extent to which students were acquiring such general hypothesis-testing skills. The quizzes were administered in lab sections following investigations in which students generated and tested the alternative hypotheses listed. Each quiz asks the same questions about testing alternative hypotheses.

More specifically, each quiz was designed to assess the extent to which students could generate if–and–then–therefore hypothetico-deductive arguments, complete with evidence, that would allow rejection of the alternative hypotheses. For example, consider the following argument and evidence that leads to the rejection of the weight hypothesis on the Pendulum quiz—a quiz patterned after Inhelder and Piaget's classic pendulum task (Inhelder & Piaget, 1958, pp. 67–79):

> *If* . . . differences in swing speeds are caused by differences in the amount of weight hanging on pendulums (weight hypothesis)
> *and* . . . the weights of two pendulums are varied, while holding other possible causes constant (proposed test)
> *then* . . . pendulum swing speed should vary (predicted result).
> *But* . . . suppose that the proposed test is actually carried out and the swing speed does not vary (observed result).
> *Therefore* . . . we would conclude that differences in swing speeds are probably not caused by weight differences, i.e., the weight hypothesis is probably wrong (conclusion).

Table 1
The quizzes

Pendulum quiz	
	A swinging string with a weight on the end is called a pendulum. What causes pendulums to swing fast or slow?
Hypothesis 1:	A change in the amount of weight hanging on the end of the string will cause a difference in the swing speed—the lighter the weight, the faster the swing.
Hypothesis 2:	A change in the length of string will cause a difference in the swing speed—the shorter the string, the faster the swing.
Question:	How could you test these hypotheses? (a) Describe your experiment. (b) What are the predicted results of your experiment (assuming that the hypotheses are correct)? (c) What result would show that Hypothesis 1 is probably wrong? (d) What result would show that Hypothesis 2 is probably wrong?
Mealworm quiz	
	A student recently placed some mealworms in a rectangular box to observe their behavior. She noticed that the mealworms tended to group at the right end of the box. She also noticed that the right end had some leaves in it and that the box was darker at that end. She wondered what caused them to group at the right end.
Hypothesis 1:	They went to the right end because it had leaves in it.
Hypothesis 2:	They went to the right end because it was darker than the left end.
Question:	How could you test these hypotheses? (a) Describe your experiment. (b) What are the predicted results (assuming that the hypotheses are correct)? (c) What result would show that Hypothesis 1 is probably wrong? (d) What result would show that Hypothesis 2 is probably wrong?
“A” Mountain quiz	
	A recent survey of organisms on “A” Mountain revealed more grass on the north-facing slope than on the south-facing slope. In response to the causal question, “Why is there more grass on the north-facing slope?” a student generated the following hypotheses:
Hypothesis 1:	Lack of moisture in the soil on the south-facing slope keeps grass from growing there (i.e., north is better shaded from the sun’s drying rays).
Hypothesis 2:	The sunlight itself is too intense for good grass growth on the south-facing slope (i.e., very intense rays disrupt the grass’s ability to conduct photosynthesis).
Question:	How could you test these hypotheses? (a) Describe you experiment(s). (b) What are the predicted results of your experiment(s) assuming that the hypotheses are correct? (c) What result would show that Hypothesis 1 is probably wrong? (d) What result would show that Hypothesis 2 is probably wrong?
Osmosis quiz	
	When a thin slice of red onion cells is bathed in saltwater, the red portion of each cell appears to shrink. What causes the red portion to appear to shrink?
Hypothesis 1:	Salt ions (i.e., Na^+ and Cl^-) enter the space between the cell wall and the cell membrane and push on the cell membrane.
Hypothesis 2:	Water molecules (i.e., H_2O) are charged (i.e., thus leave the cell owing to attractive forces of the salt ions.
Question:	How could you use model cells made of dialysis tubing, a weighing device, and solutions such as saltwater, distilled water, and glucose to test these hypotheses? (a) Describe your experiment. (b) What are the predicted results assuming that the hypotheses are correct? (c) What result would show that Hypothesis 1 is probably wrong? (d) What result would show that Hypothesis 2 is probably wrong? (Note: These hypotheses were not intended to be scientifically valid in the sense that when tested they would be supported. Rather, the intent of the quiz is to discover if students can devise tests of the hypotheses regardless of their empirical status.)

Importantly, the quizzes were administered in the order listed. This means that if students were in fact acquiring general hypothesis-testing skills during the semester, performance should improve from quiz to quiz. Consequently, we were surprised to find that most students responded successfully on the Pendulum quiz (94%), whereas success on the Mealworm quiz dropped to 82%. Performance dropped even further to 57% on the "A" Mountain quiz and to a dismal 18% on the Osmosis quiz. What might be the cause or causes of this unexpected drop in performance?

The working hypothesis we wished to test by the present research is that the extent to which students successfully generate and test alternative hypotheses depends on the presence or absence of two levels of general hypothesis-testing skills. The first hypothesized level involves hypothesis testing in contexts in which the tentative causal agents can be directly observed or sensed or measured (e.g., the long or short strings and heavy or light weights on pendulums, the number of smelly leaves and light or dark areas at the ends of boxes), whereas the second involves hypothesis-testing in contexts in which the tentative causal agents are unobservable (i.e., imaginary or abstract or theoretical) such as Na^+ and CL^- ions and charged H_2O molecules. Thus, successful performance depends in part on the abstractness of the hypotheses in question.

To clarify this distinction between observable and unobservable causal agents, consider the nature of water. Students can observe water directly. At room temperature, water appears as a clear liquid. Thus, it is not to hard to imagine that the presence or absence of this clear liquid might influence mealworm behavior. In other words, students should have little difficulty in understanding (i.e., assimilating or representing) the hypothesis that mealworms may have moved to the right end of a box because of the clear liquid (called water) at that end. On the other hand, in the Osmosis quiz, water is no longer treated as merely as a clear liquid. Instead, it is conceived of as consisting of charged H_2O molecules. Of course, students cannot see individual water molecules to know whether each really consists of two hydrogen and one oxygen atom, much less whether each is charged. Thus, the hypothesis that unseen Na^+ and Cl^- ions leave cells because of their attraction to unseen charged H_2O molecules should be more difficult to assimilate and represent.

Increased difficulty in representing and reasoning about unobservable entities may also stem from the increased complexity of the arguments needed to test their hypothesized role(s). For example, suppose the hypothesis is generated that red onion cells shrink when bathed in saltwater because H_2O molecules exit the cells. Furthermore, suppose this molecules-exiting hypothesis is pitted against an alternative that claims that the cells just appear smaller because Na^+ and Cl^- ions push on their cell membranes (Table 1). The following argument and experiment using dialysis bags, which are assumed to have properties similar to those of cell membranes, can be used to test these alternatives:

> *If* . . . cells shrink because unobservable Na^+ and Cl^- ions push on their cell membranes (ion-push hypothesis)
> *and* . . . a dialysis bag filled with distilled water is weighed, bathed in saltwater for several minutes, and then reweighed (proposed test)
> *then* . . . the bag should appear smaller while in the saltwater, but should not lose weight (predicted result). The bag should not lose weight because the H_2O molecules, which presumably weigh some measurable amount, should not leave the bag (theoretical rationale).
> *But* . . . suppose upon conducting the experiment, we find that the bag does lose weight (observed result).
> *Therefore* . . . we would conclude that the ion-push hypothesis is probably wrong (conclusion).

Although the previous argument follows the same if–and–then–therefore form used to test hypotheses about why pendulums swing fast or slow, it also includes a theoretical rationale. The theoretical rationale is needed to link the experiment's design with the hypothesized cause(s). No such theoretical rationale is needed in the pendulum context because there the hypothesized cause and the experiment's independent variable are one and the same (i.e., the amount of weight hanging on the string).

Thus, inherent in this argument is the notion that hypothesis testing can be undertaken on two qualitatively different levels with success at testing hypotheses involving observable causal agents as a prerequisite for becoming proficient at testing hypotheses involving unobservable theoretical entities. In other words, students first become generally skilled at testing hypotheses about observable causal agents (skills that appear comparable to those of Piaget's formal operational thinker (e.g., Inhelder & Piaget, 1958; Lawson & Renner, 1975). Only then, given the necessary developmental conditions, do they become generally skilled at testing hypotheses about unobservable causal agents. Importantly, this developmental view does not claim that declarative knowledge is not needed. Rather, it is viewed as a necessary but insufficient condition for hypothesis testing. Interestingly, some hypothesis-testing situations appear to involve causal claims that fall between the extremes. For example, the second hypothesis advanced to explain the lack of grass on the south-facing slope on the "A" Mountain quiz involves very intense sunlight—a readily observable factor—but also involves grass's ability to conduct photosynthesis–a clearly unobservable process, which in theory involves unobservable entities such as CO_2 molecules, photons, electrons, and the like.

Given the distinction we are trying to make between the observable and the unobservable, one might wonder what effect technological advances such as the invention of increasingly powerful electron microscopes have had on the status of concepts such as atoms and molecules. For example, does the fact that photographs now exist presumably showing individual atoms reduce the status of the atom concept from theoretical to concrete? We think not—primarily because the photographs merely reveal images that look like little round balls. Thus, one still does not actually see atoms. In other words, deciding whether the photographs actually show atoms is still a matter of interpretation, not observation.

In summary, like William Perry's search for patterns of intellectual development during the college years (Perry, 1970), as well as those of other developmentally based researchers who have sought developmental advances beyond Piaget's formal stage (e.g., Arlin, 1975; Commons, Richards, & Armon, 1984; Epstein, 1986; Kramer, 1983; Hudspeth & Pribrum, 1990; Thatcher, 1991; Thatcher, Walker, & Guidice, 1987; Riegel, 1973), the present hypothesis attempts to understand the cognition of college students not only in terms of the amounts of declarative knowledge that they may have acquired, but by general abilities to process information and construct concepts in qualitatively more powerful ways.

Of course, an important alternative hypothesis exists that would explain the performance differences in terms of the presence or absence of declarative knowledge (as opposed to procedural knowledge) (Anderson, 1980) specific to each task (cf. Korthagen & Lagerwerf, 1995; Van Heile, 1986). In other words, according to this domain-specific knowledge hypothesis, if students have acquired the necessary declarative knowledge, they will successfully test alternative hypotheses. Lacking that knowledge, however, they will fail. The acquisition of declarative knowledge is not only a necessary condition for reasoning, it is also sufficient. In the present context, this would mean that our students had more specific declarative knowledge about pendulums and mealworms, less about grass growth on "A" Mountain, and still less about onion cells, salt ions, and H_2O molecules.

Method

Sample

The sample consisted of 667 undergraduate students (nonscience majors) enrolled in a course entitled The Living World taught at a major southwestern university during the fall semester of 1997. The students ranged in age from 15.8 to 47.1 years [mean age 19.64 years, standard deviation (*SD*) 3.02].

Design

The first step in testing the study's working hypothesis was the selection of a valid measure of scientific reasoning that included items testing students' ability to test alternative hypotheses involving observable causal agents. Lawson's Classroom Test of Scientific Reasoning was selected for this purpose (Lawson, 1978). Because the original test does not include items explicitly assessing students' hypothesis testing skills in contexts in which the hypotheses involve unobservable entities, two new items (burning candle and red blood cells) that did so were invented and added. Thus, each new item should require developmentally more advanced reasoning skills than those assessed by the original test.

The modified test was then administered to students enrolled in The Living World at the start of the Fall 1997 semester. Scores on the modified test were used to classify student responses into four categories that presumably reflected their ability to test both types of hypotheses. Students were classified into one of four categories (i.e., Level 0 = students not able to test hypotheses involving observable causal agents; Low Level 1 = students inconsistently able to test hypotheses involving observable causal agents; High Level 1 = students consistently able to test hypotheses involving observable causal agents; Level 2 = students able to test hypotheses involving unobservable causal agents). The course was then taught and records were kept of student performance on course exams. The modified test was also administered at the end of the semester to assess test-retest reliability, measure student progress in reasoning during the semester, and determine whether assessed reasoning skill at the start or at the end of the semester is the better predictor of course performance.

Next, a transfer problem that in theory required Level 2 hypothesis-testing skills was constructed and also administered at the end of the semester. The problem was considered to be a transfer problem because it was written within a context not discussed or explored in the course. More specifically, the problem involved testing a hypothesis about why balloons move forward or backward when a moving vehicle suddenly stops. Five multiple choice questions assessing the declarative knowledge thought to be needed to solve the transfer problem were also constructed and administered. Consequently, if Level 2 reasoning skills alone are sufficient to solve the transfer problem, then reasoning skill alone should predict success. On the other hand, if declarative knowledge is sufficient, then it alone should predict success. Finally, if both Level 2 reasoning skill and declarative knowledge are necessary, then both should predict success.

Because the course introduced a number of biological and biochemical theories involving both observable and unobservable causal agents, a more general prediction was also advanced: If the modified reasoning test is a valid measure of general levels of hypothesis-testing skills, then course exam scores of the Level 0 students should be significantly lower than those of the Level 1 students; and exam scores of the Level 1 students should be significantly lower than those of the Level 2 students. This prediction is based on the assumption that hypothesis-testing skills play a role in concept construction. In essence, the argument is being made that even though these sorts of exam items do not directly assess hypothesis-testing skills, such skills nevertheless play a role in construction and retention of such concepts. Presumably this is because students typically do not come to the learning situation as blank slates. Rather, they often come with alternative conceptions (i.e., hypotheses) that must be modified or replaced by scientific conceptions—thus, concept construction often engages hypothetico-deductive reasoning skills (cf. Lawson, Abraham, & Renner, 1989; Lawson & Renner, 1975; Lawson & Thompson, 1988; Lawson & Weser, 1990). On the other hand, if classification into Level 2 of the modified test does not reflect a generalizable advance in reasoning, but instead represents the acquisition of domain-specific declarative knowledge needed to respond successfully to the two new test items (i.e., the burning candle and red blood cells items), then Level 1 and Level 2 students should perform equally well.

The Course

The Living World consists of three weekly 50-min lectures (delivered by the course professor) and one weekly 2-h lab (each taught by one of 13 graduate student teaching assistants) each week for 15 weeks. In the order presented, course topics included the theories of evolution and natural selection, animal behavior theory, various physiological theories, theories of classical and molecular genetics, and theories of photosynthesis and cellular respiration. In most cases, topics were first explored and new terms first introduced in labs. Lectures then discussed the topics in more detail and applied them to additional biological and nonbiological contexts. Thus, the course employed the learning cycle method of instruction (Eakin & Karplus, 1976; Karplus, 1977; Lawson et al., 1989; Renner & Marek, 1990).

Predictor Variables

Hypothesis-Testing Skills. Hypothesis-testing skills were assessed by a 13-item written test based on reasoning patterns associated with hypothesis testing (i.e., the identification and control of variables, correlational reasoning, probabilistic reasoning, proportional reasoning, and combinatorial reasoning). As mentioned, the test was a modified version of Lawson's Classroom Test of Scientific Reasoning. With respect to hypothesis testing, the original test includes items in which the tentative causal agents are for the most part observable. For example, two items involve testing hypotheses in the context of the pendulum task mentioned above, two other items involve fruitfly responses to red and blue light, and one item involves a light bulb's response to pushed buttons.

Validity of the original test has been established by several studies (e.g., Lawson, 1978, 1979, 1980a, 1980b, 1982, 1983, 1987, 1990, 1992, 1995; Lawson & Weser, 1990; Lawson, Baker, DiDonato, Verdi, & Johnson, 1993). An important aspect of the establishment of test validity, as was the case with many of Piaget's original tasks, was the need to demonstrate that performance differences on the items were caused by differences in reasoning patterns and not by the presence or absence of domain-specific knowledge. In other words, items should require only specific knowledge that students can reasonably be presumed to have. In short, the studies have supported this presumption. The pendulum task is an excellent example of this point as all students presumably know what strings and weights are and what is meant by swinging back and forth.

The modified test used in the present study contains 11 of the original items plus two new items that are hypothesized to require Level 2 thinking skills because each requires students to use hypothetico-deductive reasoning to reject hypotheses involving unobservable entities (i.e., dissolving CO_2 molecules and pushing or attracting Na^+ and Cl^- ions). Of course, testing the validity of this claim, as opposed to the claim that the tasks merely measure the presence or absence of domain-specific declarative knowledge, is a central component of the present study. One of the items involves water rise in an inverted cylinder after the cylinder had been placed over a burning candle sitting in water. The other item involves changes in the appearance of red blood cells when bathed in saltwater. The two new items appear as follows:

The Burning Candle. Figure 1 shows a drinking glass and a burning birthday candle stuck in a small piece of clay standing in a pan of water. When the glass is turned upside down, put over the candle and placed in the water, the candle quickly goes out and the water rushes up into the glass (as shown at the right).

This observation raises an interesting question: Why does the water rush up into the glass? Here is a possible explanation: The flame converts oxygen from the air to carbon dioxide. Because oxygen does not dissolve very rapidly in water, but carbon dioxide does, the newly formed carbon dioxide dissolves rapidly in the water lowering the air pressure inside the glass. Thus, the relatively higher air pressure outside the glass pushes the water up into the glass. (a) Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). Using these materials, describe a way to test this possible explanation. (b) What result of your test would show that this explanation is probably wrong?

The Red Blood Cells. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in Figure 2, the magnified red blood cells look like little round balls. After adding a few drops of saltwater to the blood, the student noticed that the cells appeared smaller, as shown.

This observation raises an interesting question: Why do the red blood cells appear smaller? Here are two possible explanations:

1. Salt ions (i.e., Na^+ and Cl^-) push on the cell membranes and make them appear smaller.
2. Water molecules are attracted to the salt ions so water molecules move out and leave the cells smaller.

Suppose you have a beaker, some saltwater, a very accurate weighing device, and some water-filled plastic bags. Suppose the plastic behaves just like red blood cell membranes. (a) Describe an experiment using these materials to test the two explanations. (b) What result of your experiment would show that Explanation 1 is probably wrong? (c) What result of your experiment would show that Explanation 2 is probably wrong?

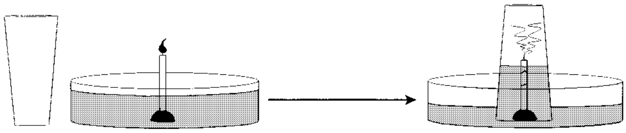


Figure 1. The burning candle.

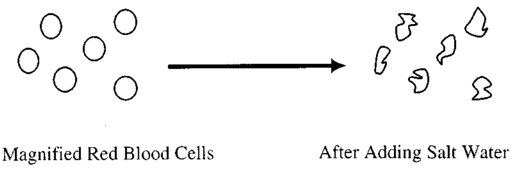


Figure 2. The red blood cells.

Scoring. All test items required students to respond to a question or make a prediction in writing and to either explain how they obtained their answer, or in the case of quantitative problems to show their calculations. Items were judged correct (a score of 1) if the correct answer plus an adequate explanation or set of calculations was present. To obtain a correct score on the burning candle item, students had to propose an adequate experiment and describe what experimental result would show that the explanation was probably wrong. Two points were possible on the red blood cells item: one for each explanation satisfactorily tested and shown to be probably wrong.

Based on the nature of the test items and the number of each type of item, scores of 0–3 were classified as Level 0 (i.e., students not able to test hypotheses involving observable causal agents). Scores of 4–6 were classified as Low Level 1 (i.e, students inconsistently able to test hypotheses involving observable causal agents). Scores of 7–10 were classified as High Level 1 (i.e., students consistently able to test hypotheses involving observable causal agents). Scores of 11–13 were classified as Level 2 (i.e., students able to test hypotheses involving unobservable entities). A test-retest reliability coefficient of .65 was obtained by comparing student performance on the test administered at the start of the semester with test performance at the semester's end.

Declarative Knowledge. The declarative knowledge believed to be involved in the Balloon Transfer problem (see below) was measured by the following multiple choice items, which were administered at the semester's end. No systematic attempt was made to introduce this knowledge during the semester.

1. Which of the following objects carries the most "umph" (momentum)?
   a. A pickup truck parked in your driveway.
   b. A pickup truck traveling at 60 miles per hour. (correct answer)
   c. A baseball traveling at 70 miles per hour.
   d. A baseball sitting on a table.
2. Air is composed of
   a. Empty space.
   b. Tiny stationary molecules.
   c. Tiny moving and colliding molecules. (correct answer)
3. Air
   a. Has weight. (correct answer)
   b. Has no weight.
4. An air-filled balloon will fall to the floor because
   a. The floor is its "natural" place.
   b. Static electricity will pull it down.
   c. It is heavier than the surrounding air. (correct answer)
   d. It is lighter than the surrounding air.

5. A helium-filled balloon will float in air because
    a. Its “natural” place is up.
    b. Static electricity will hold it up.
    c. It is heavier than the surrounding air.
    d. It is lighter than the surrounding air. (correct answer)

Scoring. Each question was scored as correct (1) or incorrect (0).

Dependent Variables

The Balloon Transfer Problem. A videotape was shown during the final laboratory period. The videotape showed a side view of a rubber balloon hanging by a string from the ceiling of a moving vehicle. Also shown was a floating mylar balloon attached by a string to the vehicle’s back seat. When the vehicle came to an abrupt stop, the hanging balloon swung forward and the floating balloon swung backward. After viewing this on the videotape, students read the following and responded in writing:

> As you could see in the video, when the vehicle stopped, the hanging balloon went forward and the floating balloon went backward. This observation raises an interesting question: Why did the hanging balloon go forward while the floating balloon went backward? Here is a possible explanation: The hanging balloon is relatively heavy; so its momentum carried it forward when the vehicle stopped. The floating balloon, being lighter than air and having less momentum, went backward because as the vehicle stopped, the heavier air molecules inside the vehicle rushed forward and piled up at the front. Thus, the piled-up air molecules at the front pushed harder on the front side of the balloon than the relatively fewer air molecules on the balloon’s backside. Thus, the balloon was pushed backward.
>
> Suppose you have two balloons just like those shown in the video, a large airtight chamber on wheels, and a vacuum pump (a pump that can extract air from airtight chambers). (a) Describe an experiment using these materials to test the possible explanation. (b) What result of your experiment would show that the explanation is probably wrong?

Scoring. Responses were judged to be correct (a score of 1) or incorrect (a score of 0). All responses were evaluated by a single rater based on the criterion that a correct response must contain the following experiment and argument: First, secure the two balloons in the chamber as they were secured in the vehicle. Next, use the pump to extract air from the chamber. Then set the chamber in motion and quickly stop it. If the balloons behave as they did in the vehicle (i.e., the hanging balloon moves forward and the floating balloon—which would now just rest on the seat—moves backward), then the explanation is probably wrong. Although this experiment and argument do not explicitly follow the if–and–then–therefore hypothetico-deductive form, it was nevertheless assumed to have been used (i.e., *If* . . . the lighter-than-air balloon went backward when the vehicle in the videotape stopped because air molecules piled up at the front and pushed it backward (molecules-push hypothesis), *and* . . . the described experiment is conducted (proposed test), *then* . . . the lighter-than-air balloon should not move backward (predicted result). The lighter-than-air balloon should not move backward because no air molecules remain in the chamber so they could not push it backward (theoretical rationale). *But* . . . suppose the proposed experiment is conducted and the lighter-than-air balloon still moves backward (actual result). *Therefore* . . . we would conclude that the explanation is probably wrong (conclusion). Interrater agreement with a subset of 100 student responses was 91%.

Lecture Examinations.

Three lecture exams written by the course professors were administered during the semester. Each exam contained 26–40 multiple choice items. Exams were machine scored with scoring adjusted so that each exam was worth 100 points for a total of 300 possible points. Table 2 contains example exam items. These items assess understanding of theoretical conceptual systems as evolution, natural selection, combustion, energy transfer and loss within food chains, population regulation, gene transfer, and reproductive strategies.

Results

Student Performance on the Study Variables

Figure 3 shows student performance on the test of hypothesis-testing skills administered at the start of the semester and again at the end of the semester. Based on pretest performance, students were classified into reasoning levels as follows: 66 students (11%) scored 0–3 and were classified at Level 0; 198 students (34%) scored 4–6 and were classified at Low Level 1; 268 students (46%) scored 7–10 and were classified at High Level 1; and 52 students (9%) scored 11–13 and were classified at Level 2.

Posttest scores improved considerably, dependent $T = 29.6$, $df = 513$, $p < .001$. Based on the same scoring criteria, numbers and percentages of students at each reasoning level on the posttest were as follows: 12 students (2%) at Level 0, 71 students (11%) at Low Level 1, 288 students (43%) at High Level 1, and 296 students (44%) at Level 2.

Declarative knowledge scores were moderately high, with 354 students (53%) responding correctly to all five questions, 221 students (33%) responding correctly to four questions, 68 students (10%) responding correctly to three questions, 13 students (2%) responding correctly to two questions, 8 students (1%) responding correctly to one question, and 3 students (<1%) responding correctly to none of the questions. The following percentages of students responded correctly to the respective questions: Question 1 = 72%; Question 2 = 95%; Question 3 = 75%; Question 4 = 95%; and Question 5 = 97%.

Overall mean score on the three lecture exams 212 points, $SD = 66.3$. This represents a 71% success rate. Success rate on the Balloon Transfer problem was 57%.

Intercorrelations among Study Variables

Table 3 shows Pearson product-moment correlation coefficients among the study variables. All coefficients were significant ($p < .01$), with the highest coefficient between the pre- and posttest hypothesis-testing skills measures (.65). The next highest coefficient was between the posttest hypothesis-testing skills measure and lecture exams (.52). The lowest coefficient was between declarative knowledge and performance on the Balloon Transfer problem (.13).

Predicting Performance on the Balloon Transfer Problem

A two-way analysis of variance was conducted in which hypothesis-testing skills level (posttest) and declarative knowledge score (0–5) were used as predictors of performance on the

Table 2

Example exam items

Which of the following is not a component of Darwin's theory of natural selection?

a. Offspring tend to resemble their parents.

b. Environments place limits on survival and reproduction.

c. Individuals with heritable traits that enhance reproductive success leave more decedents than individuals lacking those traits.

d. Within populations, lots of variation in traits can be observed among individuals.

e. None of the above. (correct answer)

The protective coloration of many insect species is a good example of

a. A vestigial trait.

b. An acquired characteristic.

c. One-step evolution.

d. An adaptation. (correct answer)

e. Speciation.

Within a habitat, which of the following organisms would be least abundant?

a. Herbivorous insects.

b. Plants.

c. Fungi.

d. Eagles. (correct answer)

e. Termites.

Scientists have concluded that increases in the concentration of CO_2 in the atmosphere over the past 100 years have been caused by

a. Decreased rates of plant photosynthesis due to lower light intensities.

b. Increased releases of volcanic gases.

c. Mobilization of long-term storage pools of carbon (fossil fuels, forests). (correct answer)

d. Increased bacterial growth in contaminated soils.

e. Atmospheric chemical reactions that release CO_2 from organic molecules.

Which of the following is a good example of a density-independent population-regulating factor?

a. Contagious disease.

b. Warfare and fighting.

c. Malnutrition.

d. Temperature drop to lethal levels. (correct answer)

In pea plants, the purple allele is dominant to the white allele. A homozygous pea plant with purple flowers is crossed with a plant with white flowers. What percentage of the offspring will have white flowers?

a. 0%. (correct answer)

b. 25%.

c. 50%.

d. 75%.

e. 100%.

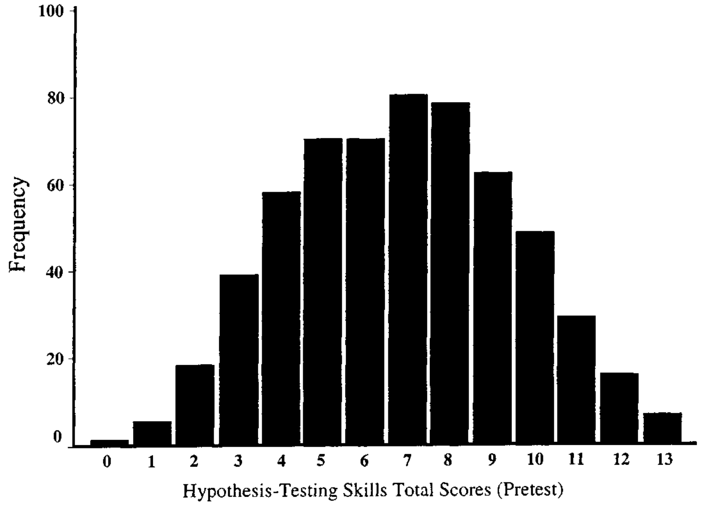
Species vary widely in their reproductive potential. Some have many offspring, whereas others have few. Which of the following species would you expect to survive better under highly competitive conditions?

a. Small egg species because they can produce more offspring so at least some would survive.

b. Small egg species because they spend less energy producing eggs so the parents have more time to ensure their own survival.

c. Large egg species because their young would start life larger and better able to compete. (correct answer)

d. Large egg species because parents would waste less time laying eggs and thus have more time to secure more mates.



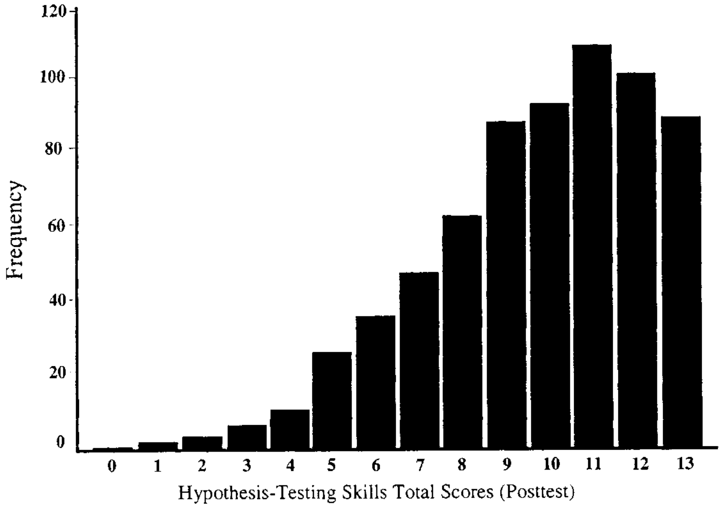


Figure 3. Frequencies of total scores on the hypothesis-testing skills test at the semester's start (pretest) and end (posttest).

Balloon Transfer problem (score of 0 or 1). The analysis revealed a significant main effect, $F_{8,657} = 5.47$, $p < .001$, a significant effect for hypothesis-testing skills level, $F_{3,663} = 8.25$, $p < .001$, and a significant effect for declarative knowledge, $F_{5,661} = 2.35$, $p < .05$.

A stepwise multiple regression analysis was conducted to determine which predictor variable (declarative knowledge or hypothesis-testing skills) was the better predictor of performance on the Balloon Transfer problem. The analysis revealed that hypothesis-testing skills, but not declarative knowledge, accounted for a significant amount of variance. However, the range of performance on the declarative knowledge measure was restricted, as 643 of the 667 students

Table 3
Pearson product-moment correlation coefficients among study variables

	Hypo. Pre.	Hypo. Post.	Decl. Know.	Exams	Balloons
Hypo. Pre.	1.00				
Hypo. Post.	0.65	1.00			
Decl. Know.	0.25	0.31	1.00		
Exams	0.36	0.52	0.26	1.00	
Balloons	0.15	0.21	0.13	0.18	1.00

Note. All p's < .01. Hypo. Pre. = hypothesis-testing skills pretest score; Hypo. Post. = hypothesis-testing skills posttest score; Decl. Know. = Declarative knowledge score; Exams = total score for the three semester exams; Balloons = score on the Balloons Transfer problem.

(96%) responded correctly to three or more of the items. Perhaps with a greater range in scores, declarative knowledge would also have been a significant predictor.

Table 4 shows relationships among hypothesis-testing skills level, declarative knowledge, and Balloon Transfer problem performance in more detail. Success on the Balloon Transfer problem improved consistently with hypothesis-testing skills level (i.e., combined column percentages are 17% success at Level 0, 33% success and Low Level 1, 57% success at High Level 1, and 65% success at Level 2). These percentages are shown graphically in Figure 4. The combined row percentages shown in the far right-hand column and in Figure 5 suggest that declarative knowledge is not as good a predictor of success on the Balloon Transfer problem. Although the respective combined row percentages of 33%, 50%, 15%, 43%, 59%, and 61% do not show a consistent increase with amount of declarative knowledge, only 24 students fell into the lowest three categories; thus, these percentages may not be representative.

Predicting Lecture Exam Scores

Figures 6 and 7 show the relationship between hypothesis-testing skills as assessed by both pretest and posttest measures and course performance as determined total scores on the three

Table 4
Relationships among hypothesis-testing levels (posttest), declarative knowledge, and Balloon Transfer problem performance: Fraction and percent correct

Declarative	Hypothesis-Testing Skills Level				Combined
Knowledge	Level 0	Low Level 1	High Level 1	Level 2	Rows
0	0/1 (0%)*		1/2 (50%)		1/3 (33%)
1		1/2 (50%)	2/5 (40%)	1/1 (100%)	4/8 (50%)
2		0/2 (0%)	1/10 (10%)	1/1 (100%)	2/13 (15%)
3	0/2 (0%)	6/20 (30%)	13/29 (45%)	10/17 (59%)	29/68 (43%)
4	1/7 (14%)	6/18 (33%)	80/120 (67%)	43/76 (57%)	130/221 (59%)
5	1/2 (50%)	11/29 (34%)	67/122 (55%)	136/201 (68%)	215/454 (61%)
Combined columns	2/12 (17%)	24/71 (34%)	164/288 (57%)	191/296 (65%)	

*Fraction and percentage of students in declarative knowledge and hypothesis-testing skill category responding correctly to the Balloon Transfer problem.

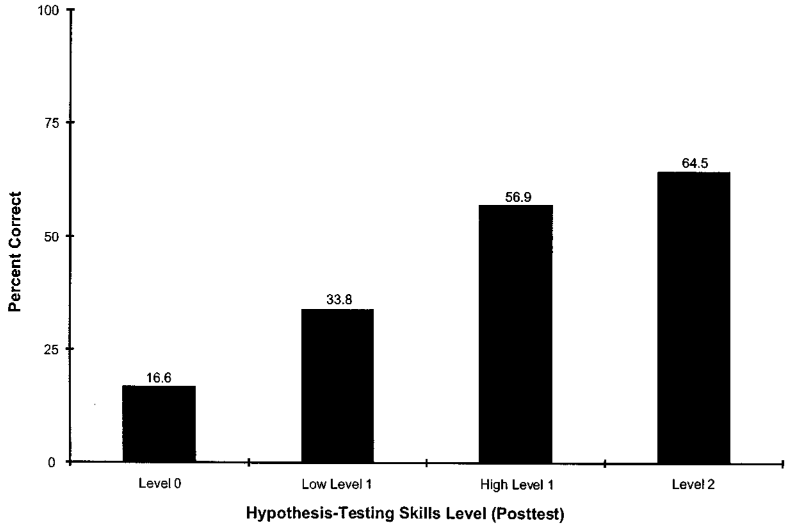


Figure 4. Relationship between hypothesis-testing skills level (posttest) and Balloon Transfer problem (percent correct).

lecture exams. The predicted relationship between hypothesis-testing skills and exam performance was found, $F_{3,583} = 2.20$, $p < .001$ for the hypothesis-testing skills pretest, and $F_{3,666} = 2.26$, $p < .001$ for the hypothesis-testing skills posttest. Tukey's post hoc tests conducted on both the pre- and posttest showed that the mean scores of all group pairs differed significantly, $p < .05$.

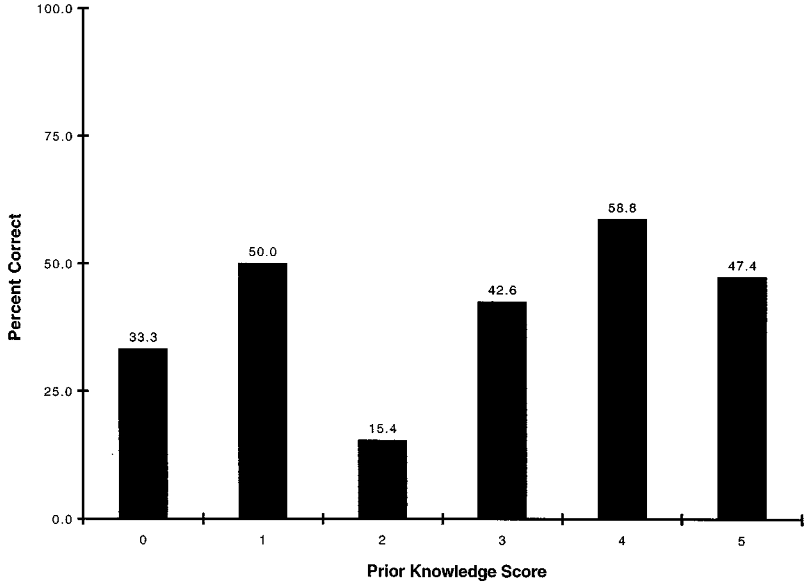


Figure 5. Relationship between declarative knowledge score and Balloon Transfer problem (percent correct).

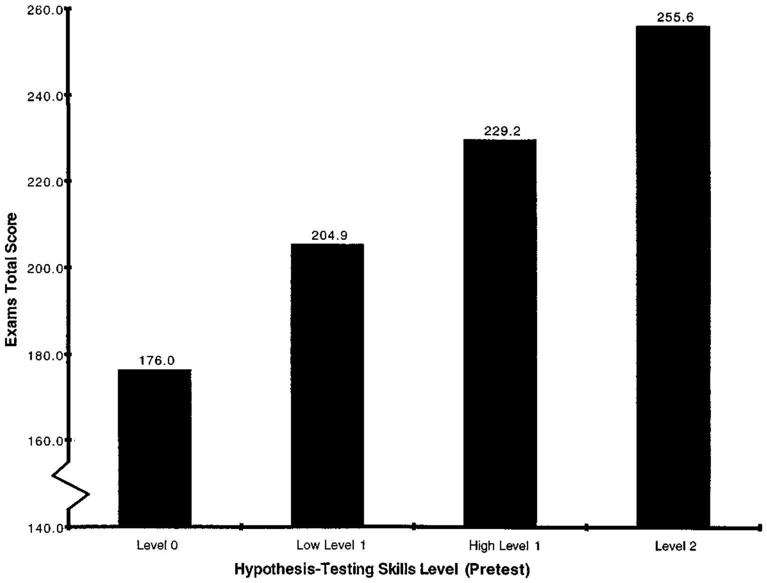


Figure 6. Relationship between levels of hypothesis-testing skills (pretest) and course performance as assessed by exams total score.

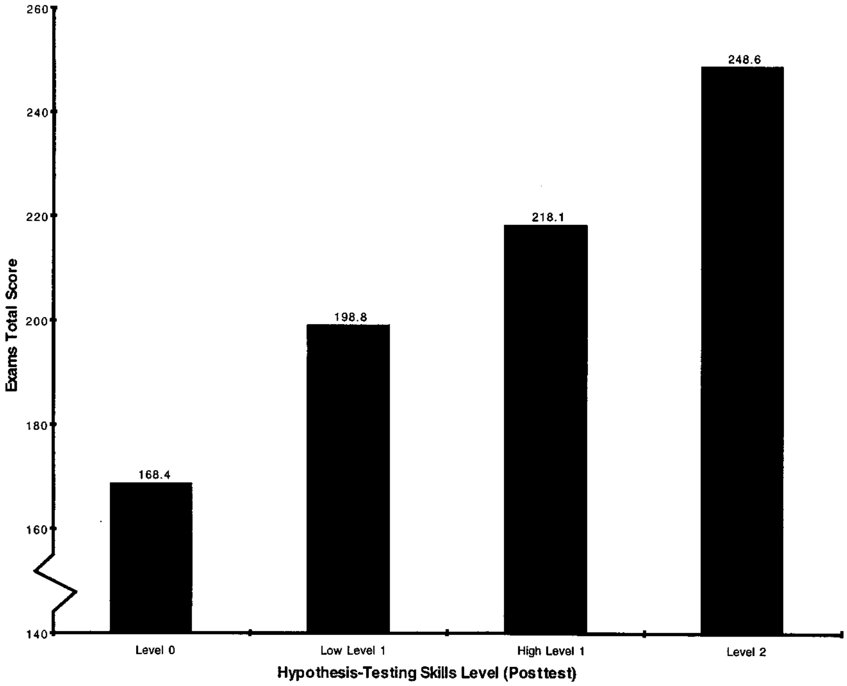


Figure 7. Relationship between levels of hypothesis-testing skills (posttest) and course performance as assessed by exams total score.

Discussion

Although the primary purpose of this study was not to test hypotheses about how to promote the development of hypothesis-testing skills, possible causes of the observed pre- to posttest gains [Figure 3] deserve mention. First, small pre- to posttest improvements have been traced to a test-retest effect (e.g., Lawson, Nordland, & DeVito, 1974). However, our former students' relatively poor performance on quizzes such as the "A" Mountain and Osmosis quiz (Table 1) strongly suggests that the substantial improvements in hypothesis testing found here are difficult to come by and not likely to have been caused by a test-retest effect. Perhaps the most reasonable explanation for the substantial gains is that the students in fact became better at testing alternative hypotheses and that this improvement came about because the course professors and graduate teaching assistants made a very conscious and concerted effort to make alternative hypothesis testing the central theme of nearly every lecture and virtually all labs. Also, because previous research on the teachability of reasoning skills suggests that hypothesis-testing skills develop best if students are given repeated opportunities to test hypotheses in familiar and observable contexts before attempting to so with unobservable entities, the labs and lectures were sequenced as such. For example, Westbrook and Rogers (1994) found that a 6-week ninth-grade unit on simple machines (e.g., levers, pulleys, inclined planes) with readily observable variables was successful in promoting Level 1 hypothesis-testing skills when students were explicitly challenged to generate and test alternative hypotheses. Also, Shayer and Adey (1993) found that the Thinking Science Program (Adey et al., 1989) was successful in boosting the achievement of students on the British National examinations not only in science and mathematics, but in English as well. The Thinking Science Program is designed to promote scientific thinking skills by exploring patterns and testing hypotheses first in observable contexts such as pitch pipes, shopping bags, and bouncing balls, and then in unobservable contexts such as dissolving and burning chemicals. In short, it appears that similar efforts in the present course paid off for many students.

Results of the initial test of the study's central working hypothesis, which involved assessing student performance on the Balloon Transfer problem, were somewhat equivocal. We argued that if Level 2 reasoning skills alone were sufficient to solve the transfer problem, reasoning skill alone should predict success. On the other hand, if declarative knowledge were sufficient, it alone should predict success. Finally, if both Level 2 reasoning skill and declarative knowledge were necessary, both should predict success. Based on results of the stepwise multiple regression analysis, it appears that hypothesis-testing skills, but not declarative knowledge, significantly accounted for performance differences on the Balloon Transfer problem (Table 4 and Figures 4 and 5). Level 2 students were in fact more successful than their less-skilled peers. However, Table 4 shows that only 65% of the Level 2 students responded successfully. One might wonder why the other 35% of the Level 2 students did not. It appears that their failure did not arise because they lacked some specific bit of knowledge because, as noted, declarative knowledge did not predict problem success very well, particularly with the influence of hypothesis-testing skills held constant. If Level 2 skills are sufficient and truly generalizable, success should have been higher. However, we would not expect 100% success. This is because, in theory at least, even for someone who knows in a general sense how to test Level 2 hypotheses (i.e., imagine some test condition that allows the deduction of a specific prediction that may in fact not happen), each hypothesis-testing context is different; thus, deciding how to test any specific Level 2 hypothesis requires an element of creativity. In other words, even if someone understands what he is supposed to do to test a hypothesis, he may not be able to come up with a good way to do so in any one context—particularly when given limited time, as was the case here. The case of physiologist Otto Loewi, who struggled for 17 years before he literally dreamed up a way to test his chemical transmission hypothesis, is a classic example of this point (Koestler, 1964, p. 205). This interpretation seems consistent with that of Perkins and Salomon (1989, p. 19) when they claimed that general cognitive skills exist but always function in contextualized ways.

Also note that 17% of the Level 0 students, 34% of the Low Level 1 students, and 57% of the High Level 1 students responded successfully to the Balloon Transfer problem. If Level 2 hypothesis-testing skills are indeed necessary, none of these students should have been successful. Perhaps the unexpected success of these students can at least partially be explained by the presence of overly suggestive hints contained in the wording of the problem (i.e., test the explanation using an airtight chamber on wheels and a pump that can extract air from airtight chambers).

As expected, Level 2 students performed significantly better than Level 1 students on the semester exams (Table 3 and Figures 6 and 7). Therefore, support has been found for the hypothesis that the skills used to test hypotheses involving unobservable entities exist and are of general use in course performance (i.e., in understanding theoretical concepts and in responding correctly to exam items about such concepts). In other words, had success on the new test items (the burning candle and the red blood cells items) required only knowledge specific to those items, students classified at Level 2 would not be expected to have been more successful than students classified at Level 1 on exams testing concept understanding in other knowledge domains.

Conclusions and Educational Implications

The present study provides some support for the hypothesis that generalizable hypothesis-testing skills beyond those assessed by typical sorts of Piagetian-based measures of advanced or formal operational reasoning exist and are used to test hypotheses about unobservable entities. The skills appear to have been used by the students who possessed them to succeed on the Balloon Transfer problem and on course exams covering a wide range of theoretical topics. Yet, the distinction between Level 1 and Level 2 reasoning does not appear very clear-cut as an element of creativity and perhaps one or more yet to be identified factors [e.g., confidence, internal locus of control, “emotional intelligence” as defined by Goleman (1995)] seem to play a role in determining the extent to which such hypothesis-testing skills may or may not generalize to other contexts.

Evidence suggests that the presence of declarative knowledge alone is not sufficient to produce successful hypothesis-testing performance at this abstract/theoretical level. This is not to say that declarative knowledge is unimportant to Level 2 performance. Nevertheless, a number of students who apparently lacked one or more pieces of what we presumed to be key declarative-knowledge concepts (i.e., momentum, relative density of gases, molecular nature of gases) gave evidence of having successfully used hypothetico-deductive reasoning to test a hypothesis about the cause of movement of two balloons on the Balloon Transfer problem.

Although future research is needed to explore the role of Level 2 hypothesis-testing skills in additional instructional contexts and to perhaps identify additional factors that play roles in determining whether and when students employ such skills, it seems reasonable to suggest that making Level 2 hypothesis-testing a central focus of college-level science instruction can be very effective, particularly when lectures and labs are sequenced to move from the observable and familiar to the unobservable and unfamiliar. Although our present sequencing of labs and lectures seems to be relatively effective, many students continued to exhibit difficulties in Level 2 hypothesis testing. These difficulties included a continued confusion between descriptive and causal questions, between hypotheses and predictions (i.e., expected results), and between evidence (i.e, observed results) and scientific conclusions. Clearly, much additional research is needed to determine how best to design curricula to eliminate these persistent problems.

This material is based on research partially supported by the National Science Foundation under Grant DUE 9453610. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation

References

Adey, P., Shayer, M., & Yates, C. (1989). Thinking science: Classroom activities in secondary science. Surrey, England: Nelson.

Anderson, J.R. (1980). Cognitive psychology and its implications. San Francisco: W.H. Freeman.

Arlin, P.K. (1975). Cognitive development in adulthood: A fifth stage? Developmental Psychology, 11, 602–606.

Baker, J.J.W., & Allen, G.E. (1977). The study of biology (3rd ed.). Menlo Park, CA: Addison-Wesley.

Burmester, M.A. (1952). Behavior involved in critical aspects of scientific thinking. Science Education, 36, 259–263.

Carey. S.S. (1998). A beginner’s guide to scientific method (2nd ed.). Belmont, CA: Wadsworth.

Cavallo, A.M.L. (1996). Meaningful learning, reasoning ability, and students’ understanding and problem solving of topics in genetics. Journal of Research in Science Teaching, 33, 625–656.

Chamberlain, T.C. (1965). The method of multiple working hypotheses. Science, 148, 754–759. (Original work published 1897)

Commons, M.L., Richards, F.A., & Armon, C. (Eds.). (1984). Beyond formal operations: Late adolescent cognitive development. New York: Praeger.

Eakin, J.R., & Karplus, R. (1976). Science Curriculum Improvement Study (SCIS) final report. Berkeley, CA: Regents of the University of California.

Epstein, H.T. (1986). Stages in human brain development. Developmental Brain Research, 30, 114–119.

Germann, P.J. (1994). Testing a model of science process skills acquisition: An interaction with parents’ education, preferred language, gender, science attitude, cognitive development, academic ability, and biology knowledge. Journal of Research in Science Teaching, 31, 749–783.

Germann, P.J., & Aram, R.J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. Journal of Research in Science Teaching, 33, 773–798.

Goleman, D. (1995). Emotional intelligence. New York: Bantam.

Hudspeth, W.J., & Pribram, K.H. (1990). Stages of brain and cognitive maturation. Journal of Educational Psychology, 82, 881–884.

Hurst, R.W., & Milkent, M.M. (1996). Facilitating successful problem solving in biology through application of skill theory. Journal of Research in Science Teaching, 33, 541–552.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic.

Johnson, M.A., & Lawson, A.E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? Journal of Research in Science Teaching, 35, 89–103.

Karplus, R. (1977). Science teaching and the development of reasoning. Journal of Research in Science Teaching, 14, 169–175.

Keys, C.W. (1994). The development of scientific reasoning skills in conjunction with collaborative assignments: An interpretive study of six ninth-grade students. Journal of Research in Science Teaching, 31, 1003–1022.

Kuhn, D. (1989). Children and adults as intuitive scientists. Psychological Review, 96, 674–689.

Koestler, A. (1964). The act of creation. London: Arkana Penguin.

Korthagen, F., & Lagerwerf, B. (1995). Levels in learning. Journal of Research in Science Teaching, 32, 1011–1038.

Kramer, D.A. (1983). Post-formal operations? A need for further conceptualization. Human Development, 26, 91–105.

Lawson, A.E. (1978). The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching, 15, 11–24.

Lawson, A.E. (1979). Relationships among performances on group-administered items of formal reasoning. Perceptual and Motor Skills, 48, 71–78.

Lawson, A.E. (1980a). The relationship among levels of intellectual development, cognitive style and grades in a college biology course. Science Education, 64, 95–102.

Lawson, A.E. (1980b). Reply to: Concurrent validity in tests of Piagetian developmental levels. Journal of Research in Science Teaching, 17, 349–350.

Lawson, A.E. (1982). The reality of general cognitive operations. Science Education, 66, 229–241.

Lawson, A.E. (1983). Predicting science achievement: The role of developmental level, disembedding ability, mental capacity, prior knowledge and beliefs. Journal of Research in Science Teaching, 20, 117–129.

Lawson, A.E. (1987). Classroom test of scientific reasoning. Unpublished manuscript, Arizona State University, Tempe, Arizona.

Lawson, A.E. (1992a). The development of reasoning among college biology students. Journal of College Science Teaching, 21, 338–344.

Lawson, A.E. (1992b). What do tests of formal reasoning actually measure? Journal of Research in Science Teaching, 29, 965–984.

Lawson, A.E. (1995). Science teaching and the development of thinking. Belmont, CA: Wadsworth.

Lawson, A.E., Abraham, M.R., & Renner, J.W. (1989). A theory of instruction: Using thelearning cycle to teach science concepts and thinking skills (NARST Monograph No. 1). Cincinnati, OH: National Association for Research in Science Teaching.

Lawson, A.E., Baker, W.P., DiDonato, L., Verdi, M.P., & Johnson, M.A. (1993). The role of physical analogies of molecular interactions and hypothetico-deductive reasoning in conceptual change. Journal of Research in Science Teaching, 30, 1073–1086.

Lawson, A.E., Nordland, F.H., & DeVito, A. (1974). Piagetian formal operational tasks: A crossover study of learning effect and reliability. Science Education, 58, 267–276.

Lawson, A.E., & Renner, J.W. (1975). Relationships of concrete and formal operational science subject matter and the developmental level of the learner. Journal of Research in Science Teaching, 12, 347–358.

Lawson, A.E., & Thompson, L.D. (1988). Formal reasoning ability and misconceptions concerning genetics and natural selection. Journal of Research in Science Teaching, 25, 733–746.

Lawson, A.E., & Weser, J. (1990). The rejection of nonscientific beliefs about life: The effects of instruction and reasoning skills. Journal of Research in Science Teaching, 27, 589–606.

Lawson, A.E., & Worsnop, W.A. (1992). Learning about evolution and rejecting a belief in special creation: Effects of reflective reasoning skill, prior knowledge, prior beliefs and religious commitment. Journal of Research in Science Teaching, 29, 143–166.

Lewis, R.W. (1988). Biology: A hypothetico-deductive science. The American Biology Teacher, 50, 362–366.

Moore, J.A. (1993). Science as a way of knowing: The foundations of modern biology. Cambridge, MA: Harvard University Press.

Noh, T., & Scharmann, L.C. (1997). Instructional influence of a molecular-level pictorial presentation of matter on students' conceptions and problem-solving ability. Journal of Research in Science Teaching, 34, 199–217.

Perkins, D.N., & Salomon, G. (1989). Are cognitive skills context-bound? Educational Researcher, 18, 16–25.

Perry, W.G., Jr. (1970). Forms of intellectual and ethical development in the college years: A scheme. New York: Holt, Rinehart, Winston.

Platt, J.R. (1964). Strong inference. Science,146, 347–353.

Renner, J.W., & Marek, E.A. (1990). An educational theory base for science teaching. Journal of Research in Science Teaching, 27, 241–246.

Riegel, K.F. (1973). Dialectic operation: The final period of cognitive development. Human Development, 16, 346–370.

Shayer, M., & Adey, P.S. (1993). Accelerating the development of formal thinking in middle and high school students. IV: Three years after a two-year intervention. Journal of Research in Science Teaching, 30, 351–366.

Thatcher, R.W., Walker, R.A., & Giudice, S. (1987). Human cerebral hemispheres develop at different rates and ages. Science, 236, 1110–1113.

Van Heile, P.M. (1986). Structure and insight, a theory of mathematics education. Orlando, FL: Academic.

Westbrook, S.L., & Rogers, L.N. (1994). Examining the development of scientific reasoning in ninth-grade physical science students. Journal of Research in Science Teaching, 31, 65–76.

Wong, E.D. (1993). Self-generated analogies as a tool for constructing and evaluating explanations of scientific phenomena. Journal of Research in Science Teaching, 30, 367–380.

Zohar, A., Weinberger, Y., & Tamir, P. (1994). The effect of biology critical thinking project on the development of critical thinking. Journal of Research in Science Teaching, 32, 183–196.