# How long does it take? A study of student acquisition of scientific abilities

Eugenia Etkina, Anna Karelina, and Maria Ruibal-Villasenor

*Graduate School of Education, Rutgers University, New Brunswick, New Jersey 08901, USA*

Most of the time, instructors of introductory physics limit their goals to students' acquisition of basic concepts and end-of-the-chapter problem solving efficiency. They overlook the development of students' science process abilities required for constructing scientific knowledge and approaching complex problems as scientists do. This goal is attainable and very valuable at the same time. This paper describes how learners improved their scientific abilities during the course of one semester and reports on the activities and facilitations that helped students in the process. We investigated how long it takes for novices to develop complex scientific abilities and whether the content and the context of the tasks affect the abilities that students demonstrate. We found that students need to conduct several cycles of scaffolded investigations to gain competence in the application of scientific abilities. Depending on the particular ability, a period of five to eight weeks of work is necessary to achieve it.

PACS number(s): 01.40.Fk, 01.40.gb, 01.50.Qb

## I. INTRODUCTION

According to recent reports in science and engineering education, students have to acquire not only conceptual and quantitative understanding of physics principles but also the science process abilities that are needed to construct explanatory models by reasoning from observations and data, to design experiments that test hypotheses, to solve complex and open-ended problems, and to work in teams with other people.[1–5] These requests place a heavy burden on the introductory physics courses for those students who will not take any more physics in college (science majors, premeds, computer science majors, etc). In addition to learning the concepts and laws of physics in a course that moves very quickly, students need to acquire abilities such as those listed above. However, instead of being one more hindrance toward a passing grade, developing scientific abilities is highly beneficial for students. They probably will not remember the details of Newton's third law or of projectile motion while treating patients or studying the effects of certain chemicals. Nevertheless, all of them will need to make decisions based on evidence, use evidence to test alternative explanations, deal with complex problems that do not have a single right answer, and work with other people in teams. Thus we suggest that it is possible to use the context of physics to help students develop abilities that they will use later in their lives.

The development of these abilities can be a complex and often frustrating process for students.[6] They struggle with problems that do not have one sole correct approach or solution. This frustration might incline instructors who engage students in these complex activities to abandon difficult and open-ended tasks and switch to clear-cut cookbook laboratories and back-of-the-chapter problems. Therefore it is very important to document how specific activities help students develop desired abilities and to have research results that show how long it takes for the students to develop these complex abilities. Such results might encourage instructors and students to cope with frustration for longer periods of time and might increase the rate of adoption of curricula aimed at developing such abilities.

More importantly, we seek to shed some light on the learning processes that take place as students acquire complex abilities which are not gained automatically by working on routine laboratory exercises[7] but grown when individuals engage in a reflective and mindful examination of the experimental problems and their own work in the laboratory. Most of the physics education research (PER) focuses on preinstruction and postinstruction measures of learning using multiple-choice instruments. Such research, although very important, misses the details of student development and does not provide information about the dynamic of learning. Meager information about post-treatment gains does not allow us to understand how the very process of learning occurs and what affects it. We need this deep knowledge of the process because one cannot design instruction if teachers do not know how students progress from A to B in learning but only know where A and B are. We must have a nuanced understanding of the process of student learning so that we can design instruction that matches the needs of our students.[8]

This paper describes the study of how students develop scientific abilities in a course that follows the investigative science learning environment (ISLE) with fully integrated instructional design laboratories. We will investigate the following research questions: (a) How long does it take for the majority of the students to develop different scientific abilities? (b) Does the time depend on the particular ability? (c) What are the factors that might affect the level of proficiency in a particular ability demonstrated by the students?

As our research purpose was not to find if the students in ISLE laboratories do develop scientific abilities but to answer questions about the process through which the learning of scientific abilities happens and the dynamic nature of this process, we followed a microgenetic approach to our study.[9] Therefore we gathered intensive data (written laboratory reports) from individual learners each week during one full semester. We scored all of the laboratory reports (weeks 1–11) of 67 students, following their progress on the relevant abilities. In addition, in order to answer the third question, we scored student responses to the open-ended laboratory practical paper-and-pencil and experimental questions. In

Secs. II–V of the paper we describe the following: (i) ISLE laboratories, briefly; (ii) the details of student work in the laboratories; (iii) the setup for the study and the data collection; and (iv) findings and discussion.

## II. SUMMARY OF ISLE LABORATORIES

ISLE (Ref. 10) is one of the reformed college physics curricula that focuses explicitly on helping students develop some of the abilities used in the practice of science.[11] A detailed description of the ISLE curriculum and scientific abilities, including the theoretical foundation of cognitive apprenticeship and formative assessment, is provided in Refs. 10–12.

The ISLE laboratories are naturally integrated in the learning process. In laboratories students design their own experiments without cookbook instructions but with the support of special guiding questions and self-assessment rubrics.[11,13] An example of a laboratory handout is provided in Appendix A; examples of several rubrics are provided in Appendix B. An example of a student laboratory report with the comments and the rubric scores is in Appendix C. What is important about the ISLE laboratories is that students have to implement different scientific abilities, such as evaluating uncertainties and assumptions. This is not only because they complete special exercises and answer questions that have the goal to develop those abilities but most importantly because they have to solve real experimental problems. For example, the students need to determine the specific heat of an object made of an unknown material. If they conduct only one experiment, there is no way to say whether the number they obtain makes any sense since there is no "accepted value." Therefore, the students need to design a second independent experiment and then make a decision on the value of the specific heat based on the assumptions in their mathematical procedure and the experimental uncertainties in their values.

In a typical laboratory, students conduct one or two experiments. All of the experiments can be grouped into three big categories.[12] The first type is observational experiment that takes place when students have to investigate a new phenomenon that they have not yet seen in large room meetings or problem solving sessions. When students design observational experiments, they need to figure out how to collect the data suggested by the laboratory handout and how to analyze the data to find patterns. For example, they need to find a pattern between the current through and potential difference across a resistor. The second type of experiments is testing experiment that students design when they need to test a hypothesis. This hypothesis is usually based on a pattern observed in a previous laboratory experiment or it is a hypothesis that students devised in other parts of the course prior to the laboratory. Sometimes they have to test a hypothesis that "a friend has devised" (these are usually based on known student ideas from the physics education research). For example, students need to test a hypothesis that magnetic poles are electrically charged.

The third type is application experiment. This is experimental problem that requires students to design several experiments to determine the value of some physical quantity—such as the coefficient of friction between their shoe and the carpet, the elastic energy stored in a spring launcher, or the specific heat of an object made of an unknown material. The application experiments, as their name suggests, are the experiments where students have to apply one or more concepts that they already know to solve the problem. The laboratory handout scaffolding questions and the rubrics are different for these three types of experiments.

We have been working on developing these design laboratories and, at the same time, researching student learning based on the analysis of their laboratory reports[13] and on direct observations of student behaviors in the laboratories for the past five years.[14] During these years, the laboratories have undergone significant revisions, and student learning has improved. In the first studies, we reported the significant positive changes in student abilities of designing an experiment, developing a mathematical procedure, and communicating the results. The abilities of evaluating uncertainties and assumptions did not improve significantly. Based on these results, we revised the laboratories, and students improved on their ability to evaluate uncertainty.[15] Finally, the latest revisions allowed us to see significant improvements in the ability to evaluate assumptions.[16] The details of the work on the improvements of the laboratories and the resulting improvements in student learning are reported elsewhere; in this paper we will describe the study that took place after these improvements were implemented.

In the present study, the laboratory handouts, laboratory discussions, and student independent work have the following elements: (a) the learning goal of the laboratory which specifies what abilities will be the focus of that particular laboratory;[17,18] (b) the actual laboratory tasks (experimental problems) that students have to accomplish during the laboratory;[13,19] (c) process-oriented guiding questions that help the students simultaneously accomplish the tasks and develop specific abilities;[14(b),20] (d) special laboratory task-related exercises that help students practice the elements of the scientific abilities;[11] (e) laboratory reflection questions that are included in the experimental design part of the laboratory and at the end of the laboratory to help students connect the laboratory experiments to the big picture of a particular physics concept and the big picture of the process of scientific inquiry;[21,22] (f) special exercises usually done as homework that allow students to read and analyze a summary of an historical scientific development (unrelated to the physics content of the laboratory task) that illustrate the actual application of a particular ability.[19] Students have to answer questions that help them reflect on the role of the abilities under analysis. These serve as models of scientific inquiry. In addition Lab Instructors provide extensive written and oral feedback to the students after students first have to develop and use a particular ability.[23,24] Students have an opportunity to revise their laboratory reports after the feedback was provided. The inclusion of these elements in the laboratories is motivated by the recommendation in the literature and our own research on student learning in the laboratories. Examples of such tasks are given in Table I.

An example of different activities in which students engage to develop a particular ability during the semester is

TABLE I. Examples of types of questions and exercises that students do to acquire the ability.

| | Examples | Ability being targeted |
|---|---|---|
| Learning goal | The goal of this laboratory is to learn to evaluate how assumptions and uncertainties affect the value of an unknown physical quantity and to choose experimental procedures least affected by assumptions and uncertainties. | Ability to evaluate theoretical assumptions and experimental uncertainties. |
| Laboratory task | Write a brief outline of your procedure including a labeled sketch. | Ability to represent ideas in multiple ways. |
| | Make predictions of the outcome of each experiment based on both ideas. | Ability to make a prediction based on an idea under test. |
| Task exercise | How might these assumptions affect the result? Be specific. | Ability to evaluate the effects of assumptions. |
| | Considering one of the relevant assumptions, evaluate its effect on the results. For example, estimate how the normal force exerted by the floor on the shoe will change if you pull the shoe not horizontally but at an angle of 5° relative to the horizontal direction? | |
| Reflection questions | How is the motion diagram different from your written descriptions of the results of the two experiments? Which is more informative? Which is more efficient? | Ability to represent ideas in multiple ways. |
| | In which case should you use a random uncertainty instead of an instrumental uncertainty? | Ability to evaluate the experimental uncertainty. |
| | Do you think that assumptions are always qualitative or they can be evaluated quantitatively? | Ability to identify assumptions. |
| Exercise | Eugenia wants to find out how fast she walked along the hiking trail that took her 1.5 h to complete. | Ability to evaluate the experimental uncertainties. |
| | She counted the number of steps with a pedometer and got 10 000 steps. In order to calculate the distance she walked she estimated the length of her stride. She walked ten steps three times at her usual rate and measured the distance with the measuring tape (the smallest division is 1 cm) and got 754, 748, and 739 cm. Find the length of her stride. What is the length of the hiking trail? What are absolute and relative uncertainties in that measurement? | |
| Lab instructors feedback | *Student's laboratory report*: "The experiment was an attempt to disprove the prediction." | Ability to distinguish between a hypothesis and a prediction. |
| | *Lab Instructor's comment*: "You cannot disprove the prediction. You can only disprove and/or support the idea by comparing the prediction and the outcome of the experiment. | |
| | *Student's laboratory report*: "The scale has uncertainty of 0.5 g. " | Ability to evaluate experimental uncertainty. |
| | *Instructor's comment*: "You have to determine the uncertainty of the final value $\mu$. Otherwise you cannot compare the results of two experiments." | |

shown in Fig. 1. This is the ability to make a prediction of the outcome of an experiment based on the hypothesis under test.[25] To master this ability, students need to understand the difference between a hypothesis and a prediction and to be able to make predictions of the outcomes of the experiments based on the hypothesis under test, not on their prior knowledge or intuition.

When students write their laboratory reports, they use rubrics for self-assessment. For example, students use the rubrics in Appendix B (last row) to self-assess the ability to make a prediction based on a hypothesis under test. Basically, after they respond to the questions in the laboratory handout, they can read the rubrics, then go back and revise their writing.

## III. DESCRIPTION OF THE STUDY

### A. Instructional setting

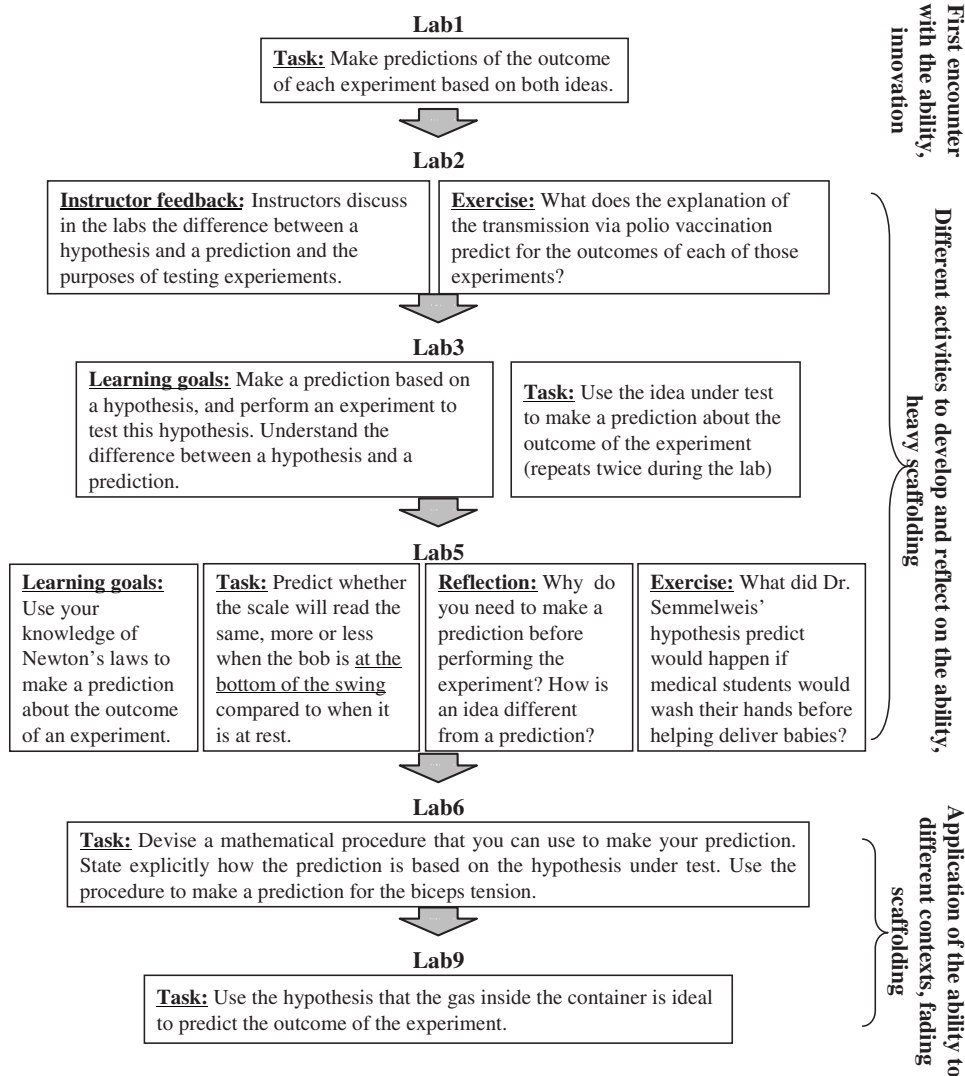The study was conducted in the first semester of a two-semester large enrollment (about 180 students) introductory

**Lab1**

**Task:** Make predictions of the outcome of each experiment based on both ideas.

↓

**Lab2**

**Instructor feedback:** Instructors discuss in the labs the difference between a hypothesis and a prediction and the purposes of testing experiements.

**Exercise:** What does the explanation of the transmission via polio vaccination predict for the outcomes of each of those experiments?

↓

**Lab3**

**Learning goals:** Make a prediction based on a hypothesis, and perform an experiment to test this hypothesis. Understand the difference between a hypothesis and a prediction.

**Task:** Use the idea under test to make a prediction about the outcome of the experiment (repeats twice during the lab)

↓

**Lab5**

**Learning goals:** Use your knowledge of Newton's laws to make a prediction about the outcome of an experiment.

**Task:** Predict whether the scale will read the same, more or less when the bob is <u>at the bottom of the swing</u> compared to when it is at rest.

**Reflection:** Why do you need to make a prediction before performing the experiment? How is an idea different from a prediction?

**Exercise:** What did Dr. Semmelweis' hypothesis predict would happen if medical students would wash their hands before helping deliver babies?

↓

**Lab6**

**Task:** Devise a mathematical procedure that you can use to make your prediction. State explicitly how the prediction is based on the hypothesis under test. Use the procedure to make a prediction for the biceps tension.

↓

**Lab9**

**Task:** Use the hypothesis that the gas inside the container is ideal to predict the outcome of the experiment.

*First encounter with the ability, innovation*

*Different activities to develop and reflect on the ability, heavy scaffolding*

*Application of the ability to different contexts, fading scaffolding*

FIG. 1. A sequence of exercises to develop the abilities (a) to distinguish between a hypothesis and a prediction and (b) to make a prediction of the outcome of the experiment based on the hypothesis.

physics course for science majors at Rutgers University in 2006. The course followed the investigative science learning environment and the laboratories were integrated. Each week students had two 55 min large room meetings, one 80 min recitation, and one 3 h laboratory. The Lab instructors in the laboratories were highly trained instructors with years of experience ranging from 2 to 8; all were members of the physics education research group. The laboratories used during the semester are available in Ref. 26. During the semester, there were 11 laboratory meetings and two laboratory practical exams.

### B. Instructional innovation

Scaffolding and support appropriate for a particular type of experiments were provided for the students in the first ten laboratories of the semester. Laboratory 11 had no scaffolding questions, no prompts, and no suggested rubrics. In addition it asked for the design of an application experiment that was based on content that was not part of the course. The laboratory task involved the consideration of drag forces in fluid dynamics. The text of the task was as follows: "Design and perform an experiment to determine the drag coef-

ficient of the helium balloon. Use this result to predict the speed of the air balloon just before it reaches the ground. Then design and perform an experiment to determine this speed. Is the result consistent with your prediction?" To emphasize the complexity of the physics involved, we repeat that the students have never before seen the terms such as Reynolds number or drag coefficient.

### C. Laboratory practical exams

There were two laboratory practical exams in the course. The first one was a paper-and-pencil exam where students had to answer six questions related to the laboratories they had worked on up until that date. Five of the questions were related to the physics of the laboratory experiments and the technical details, while the sixth was specifically related to one of the scientific abilities—the ability to test a hypothesis applied to everyday life. The question was as follows: "Describe some possible observations (related to physical, chemical, biological, or ecological phenomena of your choice). Then devise two different explanations (hypotheses that explain them). Describe what you will do to try to rule them out. For this, design testing experiments, make predic-

tions of their outcomes based on the hypotheses, and then describe the outcomes of the experiments that might make you rule out the hypotheses." The second practical exam was experimental and similar to laboratory 11 but was based on biology content. Specifically, this laboratory asked the students to determine the transpiration rate of a plant, and obviously, it was not related to the topics covered by the course. In addition, students had to use equipment, such as humidity meters, which they have not used before in the course. We want to stress that in the course, students had not learned anything about humidity, and they had not seen any humidity meters. The text of the task was as follows: "Design two experiments to determine the transpiration rate using stem cuttings from a single species of plant. Available equipment: water, beaker holding plant cuttings, parafilm, tubing, ring stand, graduated pipette, timers, humidity sensor, cup, cup with hole, scissors, and two droppers

The handout for this laboratory contained no scaffolding questions or instructions; however it provided definitions of transpiration and humidity and also included a table with saturated vapor density of water as a function of temperature. Students were not reminded to use the rubrics.

### D. Data collection

The study focused on student development of the following scientific abilities: (1) ability to identify experimental uncertainties and evaluate their effects on the result; (2) ability to minimize experimental uncertainties; (3) ability to identify assumptions made in a mathematical procedure; (4) ability to evaluate the effects of assumptions and to validate them; (5) ability to make a judgment about the results of the experimental investigation; and (6) ability to make a prediction of the outcome of the experiment based on the hypothesis under test.

For the study we collected the following data:

(1) Time distribution of the laboratory, discussion, and homework activities.

(2) Student rubric scores on seven abilities during the regular laboratories (1–10).

(3) Student rubric scores for laboratory 11 (new physics content and no scaffolding).

(4) Student rubric scores on the paper-and-pencil question related to one of the abilities that was given during laboratory practical 1.

(5) Student rubric scores on the experimental question (biology laboratory) given during laboratory practical 2.

To ensure that the rubric-based scores for student laboratory reports and practicals were reliable, we used the following procedure. For each laboratory, three trained scorers independently scored 2–3 students' laboratory reports using the chosen rubrics. Then they discussed the discrepancies in the scores to make sure that the particulars of each individual laboratory were taken into account. Then they scored an additional 7–10 randomly chosen laboratory reports until they achieved an agreement on more than 85% of the given scores (actually for many laboratories the scorers achieved almost a 100% agreement after the second scoring). Then each rater scored an additional 15–17 reports. For the laboratory prac-

tical paper-and-pencil question, we used a similar procedure. Students received their grades before we scored their work for research purposes. Examples of student laboratory work and scoring using the rubrics can be found in Appendix C.

### E. Student sample

For the study we chose three laboratory sections. The number was determined based on two considerations. (a) Each laboratory section was about 20–22 students. Each student wrote an individual laboratory report which was about three to six pages long. Realistically we had to read, score, and achieve reliability on about 3000 pages of written work. (b) There were three instructors teaching the laboratories and one instructor taught two laboratories. Thus it was reasonable to use one laboratory section per instructor. The size of the sample was 67 students. To assure that the student sample was a good indicator of the population, we administered Lawson's test of scientific reasoning as a pretest.[27] Sample students' scores on the Lawson pretest were statistically the same as the scores of the whole class. The average score of students in the selected laboratory sections is $58\% \pm 20\%$. The average score of the class is $57\% \pm 20\%$. Therefore the students in the sample represent the whole class.

### IV. FINDINGS

We present our findings as they relate to the research questions.

Question 1: *How long does it take for the majority of the students to develop different scientific abilities?*

Figures 2–7 show student progress on the development of the scientific abilities and simultaneously indicate the different activities in which students engaged in each laboratory. From the account of the activities we can say that students had multiple opportunities to develop each ability in several laboratories. We chose six abilities to trace their development (the rubrics are given in Appendix B). The results of scoring student laboratory reports are presented in Figs. 2–7. Although the figures show all the data, in order to answer the first two research questions, we will focus only on the results for weeks 1–10.

According to the scientific ability rubrics, which we use to analyze students' written reports, the report about each experiment can have a score value from 0 to 3 for each scientific ability. But in order to simplify the information presented and to make the results easier to understand, in Figs. 2–6 we have counted together the students who received scores of 0 and 1 and the students who received scores of 2 and 3 on a particular ability. Thus, in Figs. 2–6 the horizontal bars of darker shade represent the percentage of laboratory reports in the sample that received scores of 2 or 3 (that is, those that show a relative mastery of the ability) and the lighter shade bars represent the percentage of reports that received scores of 0–1 (these reports either do not show evidence of the intent to implement the ability or do not reflect any mastery). Basically, the longer the dark bar on the figure, the higher the percentage of students who achieved some mastery of the ability.
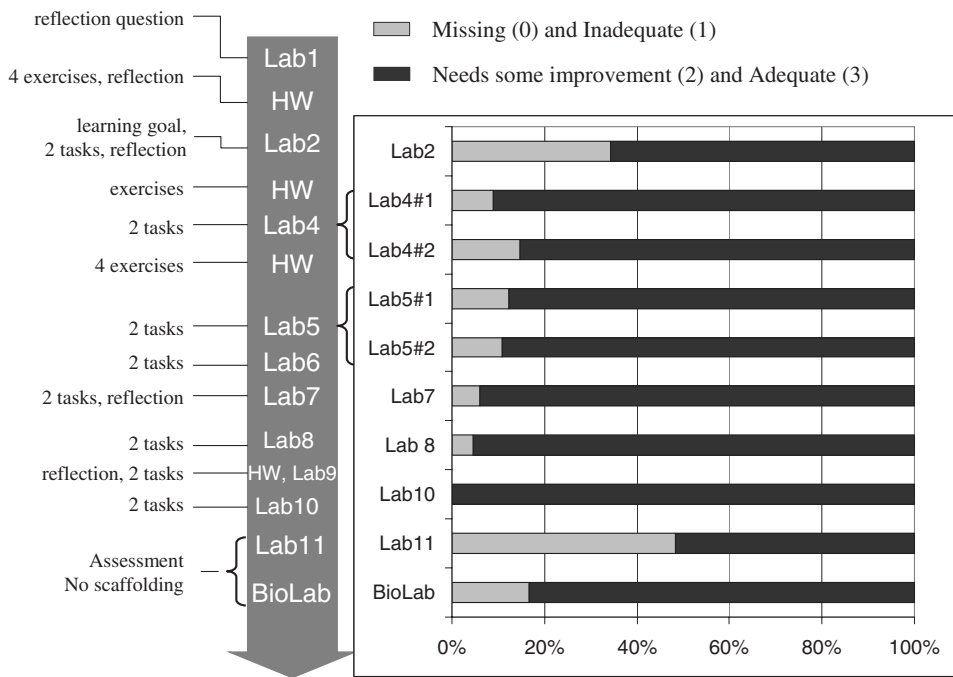
FIG. 2. Ability to identify experimental uncertainty. Laboratories 4 and 5 had two experiments each, which were scored separately. The arrow on the left shows activities related to this ability in laboratories and homework. The shaded bars represent the percentage of the students whose laboratory reports received missing and inadequate and need some improvement and adequate.

Figure 7 represents the data on how well students can distinguish between a hypotheses and predictions. For this figure we have counted together all the reports that received the scores of 0, 1, or 2 together, as the score 2 (the prediction does not describe the outcome of the experiment) indicates a serious deficiency of the report, which is more than a minor mistake or omission. Thus in Fig. 7 the dark bar represents only the students who received a score of 3. Finally, the wide vertical time bar pointing down shows what the activities were those students carried out during the laboratories or for their laboratory homework.

*Uncertainties.* Figure 2 indicates that, in the second labo-

ratory of the semester, 60% of the students were able to identify the sources of the uncertainty in their measurements and calculated values. This might look like a high number but we should not forget that the laboratories at the beginning of the semester (including laboratory 1 that was not scored with the rubrics) had many prompts to guide students in determining what brought about the uncertainty in the values and that in the second laboratory students used only a ruler and a watch. As the semester progressed, students improved on this ability so much that during regular laboratory 10 all students received scores of 2 and 3. However, evaluating uncertainty (Fig. 3), specifically determining its value
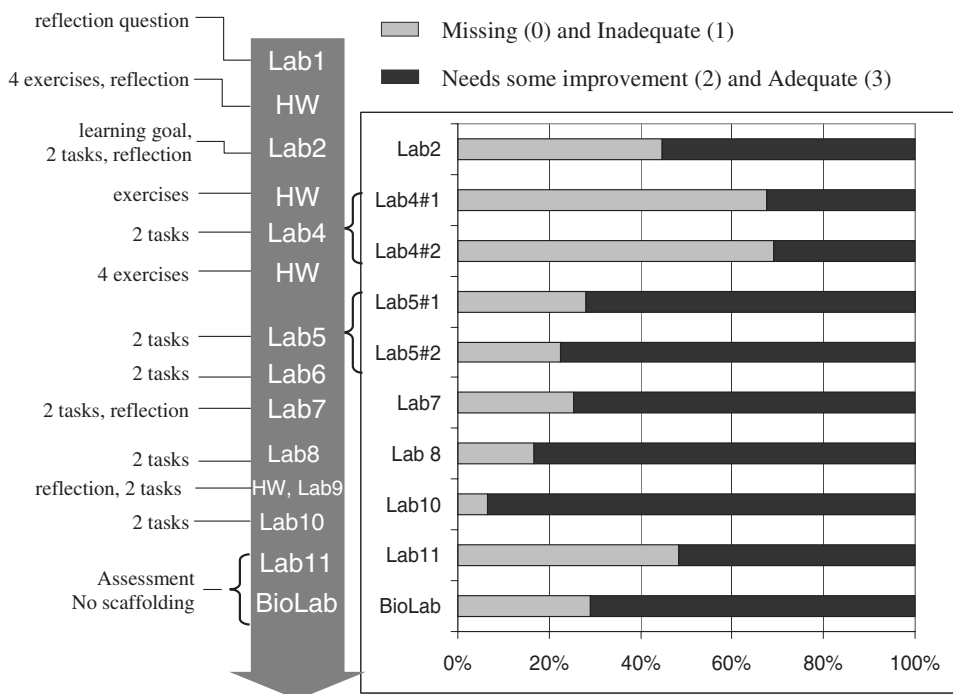


FIG. 3. Ability to evaluate experimental uncertainty. Laboratories 4 and 5 had two experiments each, scored separately. The shaded bars represent the percentage of the students whose laboratory reports received scores shown at the top of the figure.
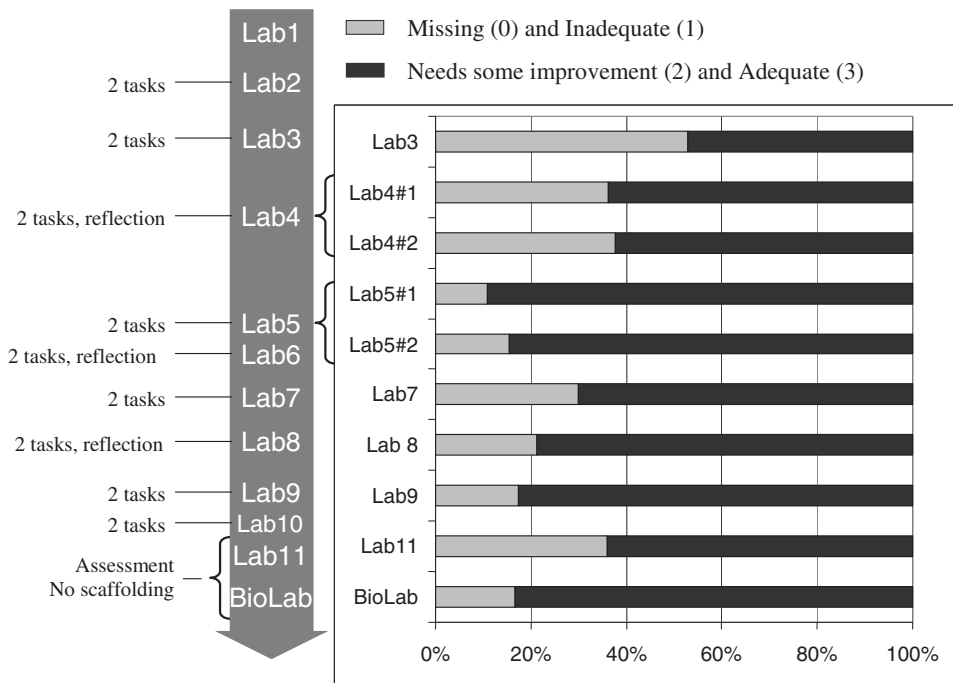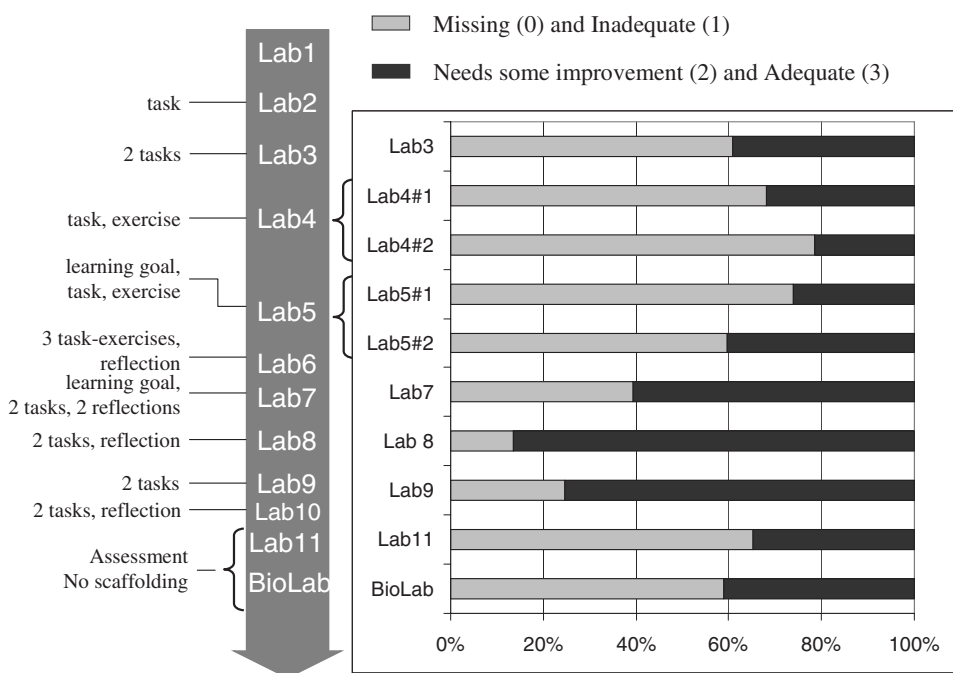
FIG. 4. Ability to identify assumptions. Laboratories 4 and 5 had two experiments each, scored separately. The shaded bars represent the percentage of the students whose laboratory reports received scores shown at the top of the figure.

by writing the result as an interval, turned out to be a much more difficult ability to acquire.[28] Students' performance grows steadily; however, by the end of the semester they achieve almost the same level as on the previous ability.

Notice how the scores of 2 and 3 drop for laboratory 4 and then increase rapidly. We believe that this can be explained by the fact that in laboratory 4 students had to design two independent experiments to determine the maximum coefficient of static friction between their shoes and the flooring. The task presented for them a considerable challenge, and it is possible that they just did not have the time to evaluate uncertainty carefully. Another explanation is that they ran out of "steam." In laboratory 5 they had to design

two different experiments to determine the net force exerted on the bob of the conical pendulum while it is in motion. The scores are much higher there possibly because the experiments were less laborious and students got used to such tasks. Overall, the results show that the number of the students who could write the result as an interval instead of just one number almost doubled over the course of the semester. The final percentage of students who mastered the ability at the level of 2 or 3 is almost 90%.

*Assumptions.* Figure 4 shows the student ability to identify assumptions in the mathematical procedure that they used. (Notice that our first scoring results are for laboratory 3, although students had two tasks in laboratory 2. Labora-



FIG. 5. Ability to evaluate the effects of assumptions. Laboratories 4 and 5 had two experiments each, scored separately. The shaded bars represent the percentage of the students whose laboratory reports received scores shown at the top of the figure.
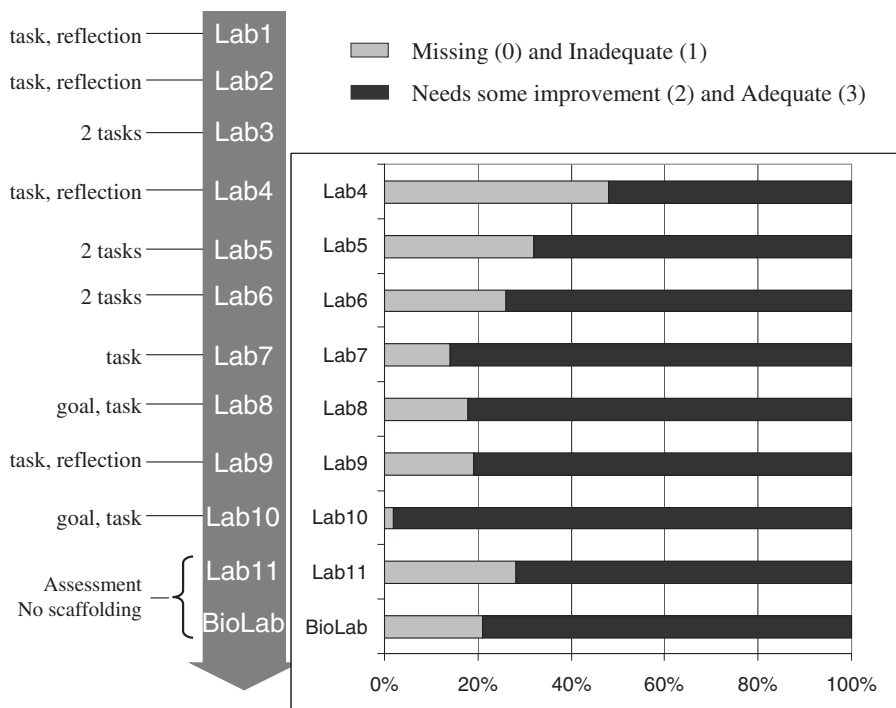
FIG. 6. Ability to compare two values and to make a judgment about the results of the experiment. The shaded bars represent the percentage of the students whose laboratory reports received scores shown at the top of the figure.

tory 2 was the longest laboratory of the semester; the students could barely finish the experiments, mostly focusing on the uncertainties that were crucial for them to solve the problem. Thus we chose not to score it for assumptions.) Here, again, we see the scores almost double by laboratory 5 and then the scores go down slightly in laboratory 7—which again had complex experiments. It appears that after week 5, students oscillate around 70%–80% on this ability. The ability to evaluate the effects of assumptions (Fig. 5) appears to be a much more difficult ability than to just identify the assumptions. Students continue improving this ability at the end of the semester.

*Forming judgments.* Figure 6 shows that the ability to compare the results of the two experiments and decide whether they are the same or different within the experimental uncertainty improves steadily during the semester. We can say that it saturates by about week 7 when 80% of the students become proficient in it. The increase in laboratory 10 is due to the specifics of the task where students had to repeat

the experiment if they could not explain the discrepancy of the results after accounting for uncertainties and assumptions.

*Making predictions.* Differentiating between hypotheses and predictions and making a prediction of the outcome of the experiment based on the hypothesis under test are difficult abilities. The percentage of students that demonstrated the ability oscillated around 60%. We need to keep in mind that in this figure (Fig. 7), the dark bar represents only those who got the score of 3, the highest level of mastery as assessed by the rubrics. Possibly, if there were more tasks for the students, their improvement would be even higher.

Question 2: *Does the time needed for the students to develop scientific abilities depend on the particular ability?*

As we have described in Sec. III, the time that it takes for the students to demonstrate mastery in the exercise of scientific abilities depends on the particular ability (see Figs. 2–7). On average, most students need a time interval of around seven weeks to develop the majority of the abilities at an
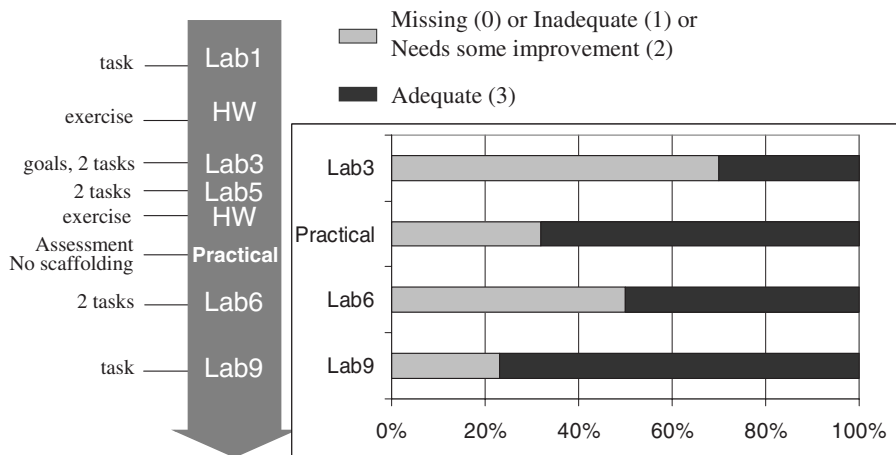


FIG. 7. Ability to distinguish between hypotheses and predictions. The dark shaded bar represents the percentage of the students whose laboratory reports received the score of 3 (adequate).

TABLE II. Correlations between the accumulated scientific ability scores and the scores on the laboratories without scaffolding. Bold font shows correlations significant at level $^*p < 0.01$ or $^{**}p < 0.001$

| Scientific ability accumulated score | | Laboratory 11 | Biolaboratory |
|---|---|---|---|
| All laboratory score | Pearson correlation | 0.24 | 0.54$^{**}$ |
| | $N$ | 62 | 68 |
| Last five laboratory score | Pearson correlation | 0.39$^*$ | 0.61$^{**}$ |
| | $N$ | 62 | 68 |

acceptable level as judged by the rubrics. However, some of the abilities necessitate a longer learning time, such as the ability to evaluate uncertainty or the ability to evaluate the effects of assumptions. We observed that after a certain number of weeks, the scores no longer continue to increase at the same rate but reach a plateau; we call this phenomenon saturation. The saturation level is quite satisfactory for all the abilities; in most cases, it is situated with about 70% of students demonstrating a particular ability and sometimes even reaches 90% (as for the ability to identify assumptions). The two most difficult abilities to develop, as mentioned above, never attained this saturation. We think that these results can be explained by considering that the exercises requiring the different scientific abilities do not present the same amount of difficulty for the students and that some abilities require a longer time and deeper physics understanding for their correct application than others. The ability to evaluate the effects of assumptions is the one that required greater knowledge and effort for implementation.

Question 3: *What are the factors that might affect the level of proficiency in a particular ability demonstrated by the students?*

From the scores of laboratories 1–10 we see that the content plays a role; however, in these laboratories students did not have to teach themselves the new content in order to complete the tasks. This challenge occurred in laboratory 11. We see all the scores drop for this laboratory. However, one needs to remember that in that laboratory, not only was the content unfamiliar to the students but in addition there were no prompting questions; so when students evaluated uncertainties or assumptions, they did it spontaneously on their own. Our previously reported research showed that students who have not worked on design laboratories do not demonstrate these abilities at all in a new content area.[14(c),29] Besides, the laboratory practical 2 (biolaboratory) shows much better results for all scored abilities than laboratory 11. Possibly the task itself was less demanding, and the students had more time to think and write about the uncertainties and assumptions.

In order to determine whether there is a correspondence between student performance on the final two laboratories, which did not incorporate any scaffolding and student performance on the regular laboratories, we developed a simple algorithm to compute the "accumulated scientific ability score" for all the regular laboratories (1–10), taking into account all abilities and all assignments that were scored. To calculate this value we added all of the scores; a particular student has received during the semester and divided this result by the number of scores. By doing this, students'

scores are not affected by any missed laboratory. We computed a "late semester" composite score as well, repeating the previous calculation for the last five laboratories of the semester. The correlations are reported in Table II.

As Table II shows, the regular laboratories that aggregate scores correlate strongly with the biopractical marks and not with those for laboratory 11. We attribute the weakness of this last correlation to the greater difficulty of laboratory 11 and the corresponding drop in student performance. Interestingly both laboratory 11 and late semester composite scores and biopractical and "late" composite scores are correlated significantly. It is possible that this strengthening of the correlations, if we take into account only the second half of the semester, is due to the fact that toward the middle of the semester, most of the abilities have reached saturation. Overall, we observed that when the scaffolding is removed and the content is novel, the demonstrated competence on students' implementation of scientific abilities drops.

The effect of the context and the content together can be analyzed using the data for the first practical exam, which was a paper-and-pencil exam as opposed to an experimental exam. We scored the paper-and-pencil laboratory practical question using the rubric for the ability to distinguish between a hypothesis and a prediction (the last row in Appendix ). "Describe some possible observations (related to physical, chemical, biological, or ecological phenomena of your choice). Then devise two different explanations (hypotheses that explain them). Describe what you will do to try to rule them out. For this, design testing experiments, make predictions of their outcomes based on the hypotheses and then describe the outcomes of the experiments that might make you rule out the hypotheses." The results are presented in Fig. 7, where the second bar corresponds to the practical question. Here we can see that almost 70% of students demonstrated the ability to make a prediction based on the hypothesis in the content area of their choice. A week later in laboratory 6, only 50% of students did this when they had to apply this same ability to some physics investigation. Therefore we can conclude that mastery in using the scientific abilities requires some knowledge or familiarity with the subject matter. In conclusion, based on the data collected for laboratory 11 and both practical exams, we can say that scientific abilities are content and context dependent and also depend on the amount of scaffolding and prompts provided.

## V. DISCUSSION

The purpose of the study was to find out how long it takes for students to acquire various scientific abilities and which

factors affect how well students develop and demonstrate them in a laboratory environment where they need to design their own experiments. Although students devise their own setup and invent their own procedure, their work is scaffolded through the prompts and questions of the laboratory handouts, reflection questions, and special homework readings describing how scientists came to understand particular phenomena. In addition, they self-assess and improve their work using scientific ability rubrics.

We collected and analyzed laboratory reports and laboratory practical exam data to answer our research questions. Results presented in Figs. 2–7 show that at the beginning of the semester there is a rapid growth in particular abilities, and after a certain number of weeks, student acceptable performance on a particular ability arrives at a plateau and oscillates around 70%, reaching 90% for some of the abilities. We call this "saturation." Saturation takes different time intervals for different abilities. The "easy" ones, such as an ability to identify experimental uncertainty, can be mastered relatively quickly; more difficult abilities (such as the evaluation of uncertainty) take about five weeks and the most difficult ones (such as the evaluation of the effects of assumptions) keep improving until the very end of the semester. The ability "to evaluate the effects of assumptions" did not saturate, thus we can say that it probably takes more than one semester to develop.

Most of the research done on the acquisition of scientific abilities did not investigate how much time it takes to develop them but focused on premeasurement and postmeasurement to check the effectiveness of a particular intervention. Schauble[30] reported on a study of the development of scientific reasoning in the course of self-directed investigations where children and adults designed and conducted their own investigations to answer two proposed problems during six 40 min sessions. She found that, during the course of the study, adults as well as children improved in both their understanding of the subject domain and in their strategies for generating systematic data and for making correct inferences based on pertinent evidence. However the researcher did not report about the rates in the development of these abilities.[30]

We found that the abilities are problem dependent; after students achieve a certain level of mastery in one laboratory, they might "slip" during the next laboratory. This can be explained by the fact that the capacity for implementing most abilities is fragile, dependent on the content of the laboratory, on the amount and length of the tasks, and also on the extent of the scaffolding provided. The content dependence of scientific reasoning abilities was documented in several studies that found that even professional scientists, when facing a task out of their area of expertise, show a decrease in reasoning ability.[31] The whole structure of the laboratory might affect the results: if students do not have enough time to write a detailed laboratory report and demonstrate a particular ability, we would score them low even if they used this ability in the laboratory. We also found that some abilities are more robust than others—students continue demonstrating them even when they do a laboratory related to biology instead of physics.

Figures 2–7 suggest that different activities promote the development of different abilities. In particular, we believe that it is a combination of different instructional tasks that really work. Our previous studies showed much smaller improvements and lower overall scores for the abilities to evaluate the experimental uncertainties and the effects of assumptions. But as we continued designing and refining tasks that targeted these, we observed larger gains and higher scores. It is difficult to compare our results to the other work done in this area in PER since most of the studies that investigate student experimental abilities focus only on the uncertainty of measurement; the studies do this in the context of special assessment questions, not in the context of actual student experimental work.[32,33]

When we inspect the left part of Figs. 2–7, it becomes apparent that laboratories and homeworks did not focus equally on each of the abilities. It is not surprising that the more attention we give to a particular ability, the better the results. Unfortunately, it is difficult to incorporate all of them into instruction and teachers must prioritize; thus students master some abilities better than others. We found that when students designed an experiment in an unfamiliar area of physics with no scaffolding, the level of demonstrated abilities dropped considerably.

We propose several different explanations to account for this observation. First, the cognitive load due to the content can be so high that the students do not have time or mental resources to evaluate the effects of uncertainties and assumptions as they struggle to understand the physics and design the experiment. Second, in the absence of prompting questions and references to the rubrics for self-assessment, students might disregard these aspects of the investigation. Third, for unfamiliar content, it can be very difficult to determine the implications of the assumptions contained in any procedure. We suspect that a combination of the three contributed to the observed reversion in the abilities demonstrated. These reasons could contribute differently to the drop in different abilities. For the ability to evaluate uncertainties, the probable reason was that students did not have enough time as they had to pay more attention to other elements of the laboratory, such as understanding the physics content and designing the experiment. This explanation is supported by the fact that in laboratory 11 almost all students who mentioned uncertainties evaluated them adequately. Besides, in the biolaboratory where students had plenty of time, they performed better on all of the abilities. For the ability to evaluate the effects of assumptions, the situation is different. Even those students who remembered to mention assumptions performed poorly in evaluating their effects. Also in the biolaboratory, students did not perform much better. This fact means that many students did not master this ability or did not posses sufficient knowledge of the content area and, in either case, still needed scaffolding.

We have found that students develop the ability to identify sources of uncertainty rather quickly; however, learning how to evaluate uncertainty (in particular, estimating and reporting the values of results as intervals and not single quantities) required considerably more time. Several papers have reported that most of the students in college introductory science courses do not understand uncertainty in measurements. The majority of students memorize heuristics for evaluating uncertainty without grasping the rationale behind

them. Almost all of the students believe in the existence of a "true value," ignoring the variability in data sets. Moreover students think that measurement results in a single value and not in an interval. Most of these studies used surveys to probe students' ideas about uncertainty.[7,32–34] There is one study in which researchers observed two laboratories, analyzed corresponding students' laboratory reports, inspected the final exams, and interviewed a few students.[35] In any case, none of these studies investigated the process of acquisition of this ability of evaluating uncertainties. We also observed that the ability to compare the results of two experiments, and resolve whether they are equal or different, taking into account experimental uncertainty, improves at the same rate during the semester.

We have found that the ability to identify the assumptions implicit in the mathematical procedure reached saturation after five weeks of instruction. The ability to evaluate the effects of assumptions was the one that took more time to attain. Even at the end of the semester, students continued improving. We have not found any study on student identification or evaluation of scientific assumptions. For this reason, we believe that this present work can contribute decisively to the understanding of the learning of scientific procedures and practices.

We observed that the proportion of students that were able to make predictions of the outcomes of experiments based on the hypotheses under test fluctuated around 60%. Previous studies have shown that distinguishing between hypotheses and predictions and making predictions based on hypotheses are difficult competences. Even the scientific literature frequently fails to distinguish between hypotheses and predictions.[36] There is ample literature on "selection tasks," which are the laboratory versions of choosing the best experiment to test a hypothesis. By means of these tasks, researchers can study if people apply a falsificationist strategy, and typically only 4% of the subjects try to disprove the statement. Some scholars have defended that the failure to intend to falsify the hypothesis being tested is due to the type of open-ended and inductive tasks (and not deductive) that people face in everyday life.[37] Various studies have explored and verified the "confirmation bias" defined as the general tendency to seek information that confirms beliefs or tentative hypotheses and to not search for, or ignore, those that do not support them.[38–41]

Using the results of this study, we can formulate the following implications for instruction. First, students need to complete multiple cycles of investigative tasks in order to master any ability. The students' development of scientific abilities clearly benefits from their engagement in sequences of activities aimed at a particular ability and reflection upon their work. We attribute the much faster student improvement in this study, compared with that in our previous studies, to the fact that in the older versions, the tasks were less scaffolded and the laboratories had fewer exercises directed at the attainment of scientific abilities.[13]

The process of developing each ability starts with a laboratory task that requires that particular ability; students have to invent their own procedures and generate their own solutions.[42,43] Then they reason about what they did. Students receive extended written feedback on their laboratory

reports. At home, they read about how to incorporate that ability into their investigations and how scientists exercised that same ability in the course of their research. Also students have to look at multiple scientific fields in which the ability is pertinent. Finally students creatively exercise the ability in new tasks. Students are requested to reflect about scientific practices and their work continually at all stages of the process. This sequence is based on the model of "preparation for future learning."[44,45]

The innovation-reflection-creative application model that the sequenced materials provide allows students to first try using the ability themselves with the help of prompts and questions (scaffolding), then to think and revise what they did (self-assessment), then to read or hear about different approaches (coaching), and then to apply this ability in several laboratories where they design experiments and write reports (practice with less and finally no scaffolding).

The second instructional implication is that, since it takes about five to eight weeks for the students to achieve a relative mastery of a particular ability, instructors should not get discouraged when students "do not get it" after the first laboratory. The acquisition progress is slow, and the competence shown for most of the abilities fluctuates depending on the content and the amount of scaffolding, even after eight weeks of instruction.

The third implication for instruction is that connecting each and every one of the activities and assignments with the instructional goals is an imperative undertaking for instructors. It is very easy to neglect some of the abilities if this detailed planning is omitted. Therefore, composing a map of all the activities and the abilities that they target can certainly help identify the "underaddressed" areas.

One might question whether students would develop these abilities without all these exercises and additional support just by being engaged in nondesign laboratories without extra reflection, historical reading passages, and rubrics. The answer to this question is negative.[14(c),29] Our previous studies indicate that in design laboratories without scaffolding, students do not develop the most difficult abilities, and in nondesign laboratories, student learning of scientific abilities is even poorer.[29,46]

The study has several limitations. The first is that laboratory time is limited, restricting the extent to which students might demonstrate their mastery of a particular ability. The second is that our assessment of student abilities in this study comes only from the analysis of written laboratory reports, which are a limited source of information. An essential feature of the ISLE laboratories is that students work collaboratively in groups of three or four individuals. They share responsibilities and support each other's performance and learning, just like scientists do. For this reason, even though students wrote individual reports, it is impossible to determine the particular contributions of each member of the group.

The fourth limitation is that in both laboratory 11 and practical 2 we added content and removed scaffolding, changing two variables at a time. Therefore we cannot determine the specific effects of each of the two alterations.

Finally, we studied how students developed scientific abilities in our learning environment where many tasks,

TABLE III. Some of the scientific abilities rubrics used in the study.

| Scientific ability | Missing (0) | Inadequate (1) | Needs some improvement (2) | Adequate (3) |
|---|---|---|---|---|
| G1: Is able to identify sources of experimental uncertainty. | No attempt is made to identify experimental uncertainties. | An attempt is made to identify experimental uncertainties, but most are missing, described vaguely, or incorrect. | Most experimental uncertainties are correctly identified. | All experimental uncertainties are correctly identified. |
| G2: Is able to evaluate specifically how experimental uncertainties may affect the data. | No attempt is made to evaluate experimental uncertainties. | An attempt is made to evaluate experimental uncertainties, but most are missing, described vaguely, or incorrect. | Most experimental uncertainties are evaluated correctly, though a few contain minor errors, inconsistencies, or omissions. | All experimental uncertainties are correctly evaluated. |
| D8: Is able to identify the assumptions made in using the mathematical procedure. | No attempt is made to identify any assumptions. | An attempt is made to identify assumptions, but most are missing, described vaguely, or incorrect. | Most relevant assumptions are identified. | All relevant assumptions are identified. |
| D9: Is able to determine specifically the way in which assumptions might affect the results. | No attempt is made to determine the effects of assumptions. | An attempt is made to determine the effects of some assumptions, but most are missing, described vaguely, or incorrect. | The effects of most assumptions are determined correctly, though a few contain errors, inconsistencies, or omissions. | The effects of all assumptions are correctly determined. |
| D4: Is able to make a judgment about the results of the experiment. | No discussion is present about the results of the experiment. | A judgment is made about the results, but it is not reasonable or coherent. | An acceptable judgment is made about the results, but does not incorporate uncertainties and assumptions. | An acceptable judgment is made about the results, with clear reasoning. The effects of assumptions and experimental uncertainties are considered. |
| C3: Is able to distinguish between a hypothesis and a prediction | No prediction is made. The experiment is not treated as a testing experiment. | A prediction is made but it is identical to the hypothesis. | A prediction is made an it is distinct from the hypothesis but it does not describe the outcome of a designed experiment. | A prediction is made, is distinct from the hypothesis, and describes the outcome of the testing experiment. |

teaching strategies, goals, and exchanges are intermingled, possibly causing interaction effects between variables. At this point, we do not know whether some factors are more relevant or efficient than others or whether the arrangement that we implemented in the course described in this paper was optimal; this has to be explored in the future. There are large gaps in the literature about the acquisition and development of scientific abilities because this vital educational goal has been overlooked for many years.

## APPENDIX A: HANDOUT IN A DESIGN LABORATORY FOR ONE OF THE EXPERIMENTS DURING THE SEMESTER

The application experiment: The energy stored in the hot wheels launcher. The hot wheel car launcher has a plastic block that can be pulled back to latch at four different positions. As it is pulled back, it stretches a rubber band—a greater stretch for each of the four latching positions. Your task is to use the generalized work-energy principle to determine the elastic potential energy stored in the launcher in each of these launching positions. The *available equipment* are hot wheel car, hot wheel track, hot wheel car launcher, meter stick, motion detector, masking tape, timer, scale to measure mass, and spring scale.

Write the following in your laboratory report:

TABLE IV. Student laboratory report with commentaries and rubric scores.

| Lab report of a student group | Commentary | Rubric score |
|---|---|---|
| **Experiment 1**: For each notch, shoot car up into the air, measure the distance that it goes in the air until it starts to fall. Then we can find $U_g$ and use it to find $U_s$: $U_g = U_s$. $U_s = mgy$  | Experiment solves the problem. Communication: Explanation and justification of the method with the sketch and energy bar chart. | D2: Adequate<br><br>F1: Adequate |
| We measured mass of the car, the distance it travels into the air (shoot against the wall and mark the wall where the car reaches max) $U_s = mgy$ | Data collection: All of the chosen measurements can be made and all details about how they are done are provided and clear. | D3: Adequate |
| We assume that the car is point particle, that it shoots straight up and that we can accurately measure the vertical distance. If we cannot measure accurately our values cannot be accurate | Assumptions: Most of relevant assumptions are identified. Effect of assumptions: The effect of an assumption is mentioned but described vaguely and confused with the uncertainty. | D8: Need some improvement<br><br>D9: Inadequate |
|           $y$       $mgy$<br>1st notch  0.515m  0.141 J<br>2nd notch  0.720m  0.198 J<br>3rd notch  0.945m  0.259 J<br>4th notch  1.175m  0.322 J | Data: All important data are present, organized, and recorded clearly. Analysis: Correct mathematical procedure. | G4: Adequate<br><br>D7: Adequate |
| We estimate our measurement of the distance is accurate to ±2cm. We estimate that uncertainty from assuming car is a point particle is about ±6cm (length of the car). This uncertainty is the largest so we will ignore others. This is a relative uncertainty of 11%. | Uncertainties: Most of important uncertainties are identified. Random uncertainty is not evaluated. Evaluation of uncertainty: The final result does not incorporate uncertainty. No attempt to minimize uncertainty. | G1: Adequate<br><br>G2: Inadequate<br><br>G3: Missing |
| **Experiment 2:** Launch car horizontally let car roll on floor, mark 1m from the rear of the car and measure the time it takes the car to reach the 1m mark. Thus we can calculate the car's kinetic energy.  We measure mass of the car and time it takes. $U_s = K = \frac{1}{2} mv^2$ ; $v = x/t$ | Experiment solves the problem. Explanation and justification of the method with the sketch and energy bar chart. | D2: Adequate<br><br>F1: Adequate |
| We assumed that floor is frictionless, and that all the potential energy of rubber band is transferred to the car. If floor is not frictionless our calculated $v$ will be less than $v$ immediately after launcher | Correct assumptions.<br><br>Acceptable evaluation of the effect of an assumption. | D8: Adequate<br><br>D9: Adequate |
|       time  velocity  $K = U_s$<br>1st notch  0.87s  2.3m/s  0.074 J<br>2nd notch  0.60s  3.3m/s  0.152 J<br>3rd notch  0.40s  5.0m/s  0.350 J<br>4th notch  1.35s  5/7m/s  0.455 J | All important data are present, organized, and recorded clearly. Correct mathematical procedure. | G4: Adequate<br><br>D7: Adequate |
| Our reaction time is the largest uncertainty in this experiment (±1s, relative uncertainty of 115%) | Uncertainty evaluated incorrectly (reaction time is about 0.2s, that gives about 20% uncertainty). The final result does not incorporate uncertainty. Random uncertainty is not evaluated. No attempt to minimize uncertainty. | G2: Inadequate<br><br>G3: Missing |
| | There is no discussion about the results of the experiment. Two experiments were performed but there is no discussion about the differences in the results due to the two methods. | D4: Missing<br><br>D5: Inadequate |

(a) Start by making a rough plan for how you will solve the problem. Make sure that you use two methods to determine the energy. Write a brief outline of your procedure including a labeled sketch.

(b) In the outline of your procedure, identify the physical quantities you will measure and describe how you will measure each quantity.

(c) Construct force diagrams and energy and/or momentum bar charts wherever appropriate.

(d) Devise the mathematical procedure you will need in order to solve the problem. Decide what your assumptions are and how they might affect the outcome.

(e) Perform the experiment and record the data in an appropriate manner. Determine the energies.

(f) Use your knowledge of experimental uncertainties to estimate the range within which you know the value of each energy.

(g) Which rubrics should be used to evaluate your work? Please use them.

(h) What are the *common features* between this physics experiment and the estimation of the age of the Iceman? Make a comparison table.

## APPENDIX B: SCIENTIFIC ABILITIES RUBRICS USED IN THE STUDY

For more information, see Table III.

### APPENDIX C: EXAMPLE OF A LABORATORY REPORT

A student laboratory report for the laboratory presented in Table IV. We show the scoring using all rubrics relevant for this report, not only those used for the study. The letters in the third column indicate the rubric used for scoring: letter D indicates a set of rubrics used for the ability to design an application experiment, letter G for the set of abilities to collect and analyze data, and letter F is for the set of abilities to communicate (for more information, see Table III).

[1] R. W. Bybee and B. Fuchs, Preparing the 21st century workforce: A new reform in science and technology education, J. Res. Sci. Teach. **43**, 349 (2006).

[2] P. Chin, H. Munby, N. L. Hutchinson, J. Taylor, and F. Clark, in *Reconsidering Science Learning*, edited by E. Scanlon, P. Murphy, and J. Thomas (Routledge Falmer, London, UK, 2004), p. 118.

[3] M. Coles, in *Proceedings of the Eight Symposium of IOSTE*, edited by K. Colhoun, R. Panwar, and S. Shrum (University of Alberta, Edmonton, 1997), Vol. 1, p. 292.

[4] R. Gott, S. Duggan, and P. Johnson, What do practicing applied scientists do and what are the implications for science education?, Res. Sci. Technol. Educ. **17**, 97 (1999).

[5] P. S. Lottero-Perdue and N. W. Brickhouse, Learning on the job: The acquisition of scientific competence, Sci. Educ. **86**, 756 (2002).

[6] M. Ruibal Villasenor and E. Etkina, in *2006 Physics Education Research Conference*, AIP Conf. Proc. No. 883, edited by L. McCullough, L. Hsu, and P. Heron (AIP, Syracuse, NY, 2006), pp. 105–108.

[7] T. S. Volkwyn, S. Allie, A. Buffer, and F. Lubben, Impact of a conventional introductory laboratory course on the understanding of measurement, Phys. Rev. ST Phys. Educ. Res. **4**, 010108 (2008).

[8] P. H. Winne, Minimizing the black box problem to enhance the validity of theories about instructional effects, Instr. Sci. **11**, 13 (1982).

[9] C. A. Chinn, in *Handbook of Complementary Methods in Education Research*, edited by P. B. Elmore, G. Camilli, and J. Green (Lawrence Erlbaum Associates, Mahwah, NJ, 2006), p. 427.

[10] E. Etkina and A. Van Heuvelen, in *Research Based Reform of University Physics, (AAPT)*, edited by E. F. Redish and P. Cooney (AAPT, Compadre, 2007), http://per-central.org/per_reviews/media/volume/ISLE-2007.pdf

[11] E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).

[12] E. Etkina, A. Van Heuvelen, D. T. Brookes, and D. Mills, Role of experiments in physics instruction: A process approach, Phys. Teach. **40**, 351 (2002).

[13] E. Etkina, S. Murthy, and X. Zou, Using introductory labs to engage students in experimental design, Am. J. Phys. **74**, 979 (2006).

[14] For different aspects of our laboratory research see S. Murthy and E. Etkina, in *Physics Education Research Conference*, AIP Conf. Proc. No. 790, edited by J. Marx and S. Franklin, (AIP, Sacramento, CA, 2004), p. 133; A. Karelina and E. Etkina, Acting like a physicist: Student approach study to experimental design, Phys. Rev. ST Phys. Educ. Res. **3**, 020106 (2007); E. Etkina and S. Murthy, in *2005 Physics Education Research Conference*, AIP Conf. Proc. No. 818, edited by P. Heron, L. McCullough, and J. Marx (AIP, Salt Lake City, UT, 2005), pp. 97–100.

[15] E. Etkina, A. Van Heuvelen, D. Brookes, S. Murthy, A. Karelina, and M. Ruibal-Villasenor, in *Proceedings of the National STEM Assessment Conference*, Washington, DC, 19–21 October 2006, edited by D. Deeds, G. Donald, and B. W. Callen (National Science Foundation and Drury University, Openwater Media Inc., 2008), pp. 68–80.

[16] E. Etkina, A. Karelina and M. Ruibal-Villasenor, An Invited Targeted Poster Presentation at the PERC 2008 Meetings, Edmonton, Canada, July 2008 (unpublished).

[17] R. DeShon and R. Alexander, Goal setting effects on implicit and explicit learning of complex tasks, Org. Behav. Hum. Decis. Process **65**, 18 (1996).

[18] D. L. Schunk, Goals and self-evaluative influences during children's cognitive skill learning, Am. Educ. Res. J. **33**, 359 (1996).

[19] A. Collins, J. S. Brown, and S. E. Newman, in *Knowing, Learning and Instruction: Essays in Honor of Robert Glaser*, edited by L. B. Resnick (LEA, Hillsdale, NJ, 1989), pp. 453–494.

[20] B. J. Reiser, Scaffolding complex learning: The mechanisms of

structuring and problematizing student work, J. Learn. Sci. **13**, 273 (2004).

[21] E. Boyd and A. Fales, Reflective learning: Key to learning from experience, J. Humanist. Psychol. **23**, 99 (1983).

[22] D. Boud, R. Keogh, and D. Walker, in *Reflection: Turning Experience into Learning*, edited by D. Boud, R. Keogh, and D. Walker (Kogan Page, London, 1985), pp. 7–17.

[23] J. Hattie and H. Timperley, The power of feedback, Rev. Educ. Res. **77**, 81 (2007).

[24] D. L. Butler and P. H. Winne, Feedback and self-regulated learning: A theoretical synthesis, Rev. Educ. Res. **65**, 245 (1995).

[25] A. E. Lawson, C. B. McElrath, M. S. Burton, B. D. James, R. P. Doyle, S. L. Woodward, L. Kellerman, and J. D. Snyder, Hypothetico-deductive reasoning and concept acquisition: Testing a constructivist hypothesis, J. Res. Sci. Teach. **28**, 953 (1991).

[26] http://www.islephysics.net

[27] A. E. Lawson, The development and validation of a classroom test of formal reasoning, J. Res. Sci. Teach. **15**, 11 (1978).

[28] R. Lippmann and The Physics Education Research Group, AREA Meeting, New Orleans, 2002 (unpublished) (http://www.physics.umd.edu/perg/papers/lippmann/meta_lab.pdf).

[29] A. Karelina, E. Etkina, A. Van Heuvelen, M. Ruibal Villasenor, D. Rosengrant, and C. Hmelo-Silver, Design and Non-Design Labs: Does Transfer Occur?, in *2007 Physics Education Research Conference*, AIP Conf. Proc. No. 951, edited by L. Hsu, C. Henderson, and L. McCullough (AIP, Greensboro, NC, 2007), pp. 92–95.

[30] L. Schauble, The development of scientific reasoning in knowledge-rich context, Dev. Psychol. **32**, 102 (1996).

[31] D. Willingham, Critical thinking: Why is it so hard to teach?, Am. J. Educ. **31**, 8 (2007).

[32] R. Lippmann Kung, Teaching the concepts of measurement: An example of a concept-based laboratory course, Am. J. Phys. **73**, 771 (2005).

[33] A. Buffler, S. Allie, F. Lubben, and B. Campbell, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[34] S. M. Coelho and M.-G. Séré, Pupils' reasoning and practice during hands-on activities in the measurement phase, Res. Sci. Technol. Educ. **16**, 79 (1998).

[35] M. G. Séré, R. Journeaux, and C. Larcher, Learning the statistical analysis of measurement errors, Int. J. Sci. Educ. **15**, 427 (1993).

[36] A. E. Lawson, T. rex, the crater of doom, and the nature of scientific discovery, Sci. Educ. **13**, 155 (2004).

[37] M. Oaksford and N. Chater, A rational analysis of the selection task as optimal data selection, Psychol. Rev. **101**, 608 (1994).

[38] K. A. Carlson and J. E. Russo, Biased interpretation of evidence by mock jurors, J. Exp. Psychol., Appl. **7**, 91 (2001).

[39] E. Jonas, S. Schulz-Hardt, D. Frey, and N. Thelen, Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information, J. Pers. Soc. Psychol. **80**, 557 (2001).

[40] N. Pennington and R. Hastie, Reasoning in explanation-based decision making, Cognition **49**, 123 (1993).

[41] J. Klayman and Y.-W. Ha, Confirmation, disconfirmation, and information in hypothesis testing, Psychol. Rev. **94**, 211 (1987).

[42] D. L. Schwartz and T. Martin, Inventing to prepare for future learning: The hidden efficiency of encouraging student production in statistics instruction, Cogn. Instruct. **22**, 129 (2004).

[43] D. L. Schwartz, J. D. Bransford, and D. Sears, in *Transfer of Learning from a Modern Multidisciplinary Perspective*, edited by J. P. Mestre (Information Age, Greenwich, CT, 2005), pp. 1–52.

[44] J. D. Bransford and D. T. Schwartz, in *Review of Research in Education*, edited by A. Iran-Nejad and P. D. Pearson, (AERA, Washington, DC, 1999), Vol. 24, pp. 61–100.

[45] D. L. Schwartz and J. D. Bransford, A time for telling, Cogn. Instruct. **16**, 475 (1998).

[46] M. Ruibal-Villasenor, E. Etkina, A. Karelina, A. Van Heuvelen, D. Rosengrant, and R. Jordan, From Physics to Biology: Helping Students Attain All Terrain Knowledge, in *2007 Physics Education Research Conference*, AIP Conf. Proc. No. 951, edited by L. Hsu, C. Henderson, and L. McCullough (AIP, Greensboro, NC, 2007), pp. 96–99.