*The following article has been accepted by the American Journal of Physics. After it is published, it will be found at http://scitation.aip.org/ajp.*

# Teaching the concepts of measurement: an example of a concept-based laboratory course

Rebecca Lippmann Kung
*Fysiska Institutionen, Uppsala Universitet, Uppsala, S 75121, Sweden*

For students to successfully complete an experiment, they must have an understanding of measurement and its related uncertainty. We argue for teaching the concepts of measurement and not only the calculations. An example of a concepts-based laboratory course is given, outlining the concepts presented, the design of the laboratory time, and the laboratory tasks. The concepts are briefly described and two often-overlooked concepts, predictive versus descriptive questions and internal versus external variation, are explained. Our survey results show that the fraction of students using range and not just average when comparing two data sets approximately doubled after instruction.

## I. INTRODUCTION

Instructors have several educational objectives for the physics laboratory, the most common being to link theory to practice, to develop scientific thinking, and to develop experimental skills.[1] Because most laboratory work involves measurement, it is imperative that students learn to appropriately interpret their data if any of these goals are to be met. Unfortunately, few students (at any level) can demonstrate an understanding of the uncertainty in a measurement. Students rarely carry out multiple trials spontaneously unless they suspect a flaw in their first measurement.[2] When asked to obtain multiple trials, students tend to rely only on the arithmetic mean to report a final result and to compare data sets,[3] though if they happen to get repeated values in a data set, students may report that number as the final result.[4] In general, students in the laboratory search for the "true value" and do not consider uncertainty. Many laboratory courses teach students the mathematics of uncertainty analysis such as the arithmetic mean, standard deviation, and percent error, but students are rarely able to use these constructs to make a strong argument from their data. Even worse, using such tools without understanding may be detrimental to future development of understanding.[5]

In this paper a concept-based approach to uncertainty instruction is proposed: design the student laboratory to teach concepts first, then the mathematical constructs can follow. We begin by describing the basic concepts needed for understanding uncertainty and the laboratory activities designed to teach uncertainty. Two often overlooked uncertainty concepts and one laboratory activity are described in more detail. Finally, results from this type of laboratory instruction are discussed.

## II. CONCEPTS BEFORE CALCULATION

As early as 1972, Hewitt argued for the importance of teaching students the concepts of physics before introducing the mathematical constructs,[6] and recent research-based curricula attempt to do so.[7] Research on physics education has repeatedly shown that students lack an appropriate understanding of fundamental physics concepts, even students who can successfully solve traditional physics problems.[8] The value of teaching the concepts of physics is well known and accepted for the lecture setting, but not generally for the laboratory setting.[9] Students are typically given procedures for calculating the mathematical constructs of uncertainty and are expected to absorb these ideas by following lab manuals step by step.

Given the results of research on students' understanding of acceleration,[10] special relativity,[11] electromagnetic waves,[12] it is no surprise that step-by-step-like instruction of uncertainty leaves students with, at best, the ability to successfully calculate the average and standard deviation, but with little conceptual understanding of why, when, and how to use these constructs. One way toward a solution of this problem is to create experiences that require students to build a conceptual understanding of measurement before (or perhaps along with)[13] their calculational ability.

Another argument for teaching the concepts of measurement is that most of students taking physics at the introductory level in the U.S. are not physics majors.[14] Different fields have different methods for calculating, reporting, and comparing uncertainty. Even among physicists there is much variation in the expression of uncertainty.[15] Teaching students one way of calculating and expressing uncertainty may not be useful. However, the concepts of uncertainty are applicable in any scientific domain.

## III. A CONCEPTS-BASED LABORATORY COURSE

Delineating the important ideas of a topic can be useful for defining and explaining the topic and helpful for instructors developing curricula. We propose that the underlying ideas for understanding uncertainty can be broadly categorized as follows:[16]

- All measurements have an associated uncertainty, which should be quantified and reported.
- A calculated result has an associated uncertainty based on its dependent values.
- The design of an experiment and skill of conducting the experiment affects the uncertainty in the measurement.
- Uncertainty is used to compare results and draw conclusions.

A sequence of specific laboratory concepts designed to elucidate these ideas has been implemented in a semester-long laboratory for an algebra-based introductory physics course at the University of Maryland.[17]

## A. Sequence of measurement concepts

Our sequence does not include every measurement concept necessary for a complete introductory understanding of uncertainty. We believe that reaching such a stage would require a whole series of laboratory curriculum reforms.[18] The specific measurement concepts were chosen because most of the students were unable to show mastery of these concepts before instruction and because these concepts were judged essential for students preparing for a scientific career (and are important for a wide variety of everyday situations).[19]

*Predictive versus descriptive question.* Students should understand the difference between a question asking for a prediction (often involving probability) of what might happen and a question asking for a description of what did happen. Typically a laboratory question asks a predictive question (see Sec. IIIC).

*Techniques for measuring time.* Having students measure time primarily with stopwatches makes the measurement process explicit. Using more complicated instruments may hide the underlying uncertainty factors because students are likely to think that computer sensors make perfect measurements. By using a stopwatch students are confronted with the clear limitations of a human using such a device, and can more easily be drawn into evaluating the accuracy and precision of their time measurements.[20]

*Purposes of multiple measurements.* Many students do not appreciate the purpose of multiple trials and might expect to obtain the same number repeatedly. Before proceeding to a more complicated uncertainty analysis, students should first understand the idea that multiple trials can produce a range of numbers that provide more information than just one number.

*Using range overlap.* When comparing two sets of data, one should look at the degree of overlap between the data in the two sets. The decision of how much overlap is necessary to show agreement is a difficult one and depends on specific circumstances.[21]

*Stacking.* This technique often is used on repeatable measurements to increase precision. For example, the thickness of one piece of paper can be determined by measuring the height of a stack of 500 pieces of paper and dividing by 500. Students should be able to appraise the benefits and disadvantages of stacking.[22]

*Systematic or random mechanism.* Students should think about what causes the variation in their data. Whether the mechanism is causing a systematic or random variation should affect how they attempt to reduce its effects and how they interpret their result.

*Internal versus external variation.* Students must be able to distinguish between variation coming from the measurement process, which is external to the system, and variation caused by something internal to the system being measured. They should understand that external variation detracts from their conclusions, but internal variation is neither harmful nor beneficial (see Sec. IIIC for more detail.)

*Displaying data with representations.* Introductory students tend to report their data in tables, but frequently a graph or chart is more convincing. Different types of representations convey different information and are useful for different amounts and types of data. For example, a histogram is useless for displaying the results of five readings.

*Low probability data.* Students should consider which characteristics are necessary to allow an outlying data point to be discarded.[23]

*Minimize external variation.* There are typically several ways to reduce the uncertainty of an experiment's results, which may or may not require more resources.[24] Students should consider how to improve every aspect of their data collection.

*Range propagation.* Students should know that uncertainty in experimental data will imply uncertainty in a calculated result, and should be able to find this uncertainty.

*Predict uncertainty.* Given an experimental method, we can estimate from experience what the uncertainty in the data would likely be and can estimate the uncertainty in the results. Students should be able to carry out this estimation and use it to compare experimental methods without actual implementation.

*Generalize theory.* Typically the theory taught to students is true for a limited set of circumstances (for example, when friction effects are minimal). Often, invalidating effects emerge in the student laboratory. These are not errors, but show a need to generalize the model to include such effects.

## B. Laboratory Exercises

We chose laboratory problems to illustrate the measurement concepts described in Sec. IIIA; the level of difficulty was increased as the semester progressed. They are open-ended to promote students' engagement with the measurement concepts. Because the concepts are revisited several times over the semester in different laboratory exercises (see Fig. 1), the students have the opportunity to develop an appropriate and coherent understanding of measurement. Frequently, a theoretical answer for an ideal situation is known, but the answer for the actual laboratory situation is not known. Each laboratory exercise can be answered using several different experimental designs, encouraging students to critique different experimental methods. Teaching assistants know which concepts are critical for each laboratory activity, so they can monitor the students' progress and direct students' attention to the concepts where necessary. The exercises are described in the following in the order of course implementation.

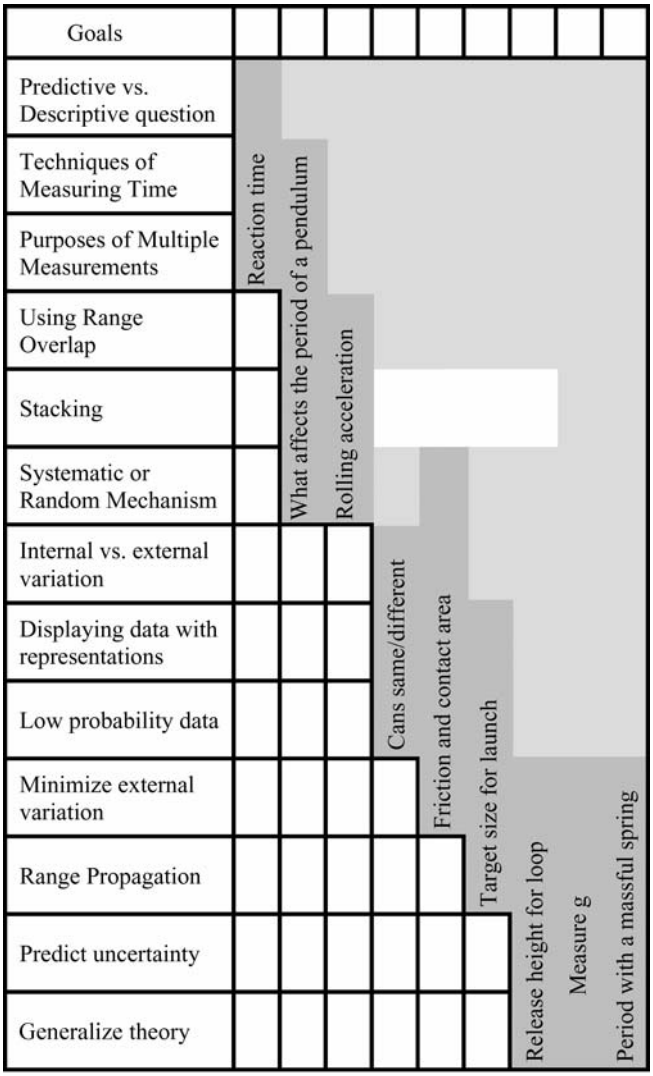


*Fig.1: The measurement goals for each laboratory question. Primary goals are shown with a darker fill than revisited goals.*

*Reaction time to catch a ruler.* In the first lab we give students a method for measuring their reaction time (the distance through which a ruler falls before it is caught) and students decide what to investigate, for example, whether a person's dominant hand (typically their right hand) has a faster reaction time. Because the class of 20 students designs the experiment, the teaching assistant can help model correct behavior and raise important issues such as how many measurements should be taken, and why.

*What affects the period of a pendulum?* Students must show whether changing the mass, length, or amplitude affects the period. Here the laboratory method is simple, allowing students time to consider questions of equipment setup, for example, whether it is better to tie the pendulum string tightly or loosely around the pivot rod, data collection, for example, whether timing 5 periods each trial is better than timing 1 period, and data analysis, for example, how much of a difference between periods is significant.

*What affects the acceleration of a rolling object?* This experiment is a typical moment of inertia laboratory. Students choose what property to investigate (for example, mass, mass distribution, length, radius, and material), and choose two objects that differ in that property. They then measure the acceleration of two objects rolling down a ramp. (Often students find that large differences in mass make a difference because of frictional effects.) Here students continue to struggle with the same concepts as the previous lab in a less familiar context.

*Are the cans the same or different?* This lab involves a collection of outwardly identical cylinders that have the same mass, but half the cylinders have a low moment of inertia and half have a high moment of inertia. Each group of students is given two cylinders and must determine whether their two cylinders' moments of inertia differ or match. Frequently the difference in students' data is caused by external variation, because students improve their measuring method with the second cylinder, and thus obtain different results. Even if students use the same method, they still must determine how large a difference is needed for the cylinders not to match. Students are told whether or not they were right, which helps them evaluate their method and analysis.

*Does friction depend on contact area?* Students measure the frictional force between a surface and an object and determine whether the force changes when the contact area changes. In this lab, small changes in method can cause large differences in result. These differences allow students to directly see the result of improving their method by minimizing external variation.

*What size target for projectile launch?* Students build a device to launch a marble off a table and determine how large a target is needed to catch the marble (see Fig. 2). To find the target size, students must determine the internal variation of their launching device, although their measurements give the external and internal variations combined.

*Release height for a ball to go around a loop.* Using energy conservation, students calculate the lowest height at which a ball should be released to complete a loop-the-loop. Students make measurements on a sample section of track to determine frictional energy losses, and thus the minimum release height for a frictional situation. They must determine how to calculate a release height from their measurements, and also decide what final value to use for the release height (for example, average, maximum, minimum, and mode).

*Measure g.* Students are asked to measure the gravitational field strength in the laboratory as accurately and precisely as possible, and to estimate the uncertainty in their answer. Here we ask students to estimate the uncertainty for several different methods before deciding on a method and taking data.

*Period of a mass oscillating on a "massful" spring.* When the mass of the spring and the mass of the oscillating object are comparable, a correction term must be added to the simple equation for the period of the oscillation. Students measure the spring constant and the period of oscilla-

tion to find the correction term, making this lab more difficult than earlier labs. Here students continue to gain experience with the same concepts in a more difficult lab.

Consider the following design problem:

> You are designing a new booth for a traveling carnival. People will use a catapult to toss a coin at a plate. If it lands on the plate, they win a prize. If no one ever wins a prize, people will stop playing. If too many people win a prize, you will lose money. So you want to make your plate large enough to catch a small fraction of coins, but not large enough to catch them all.

Two of the main things we have been working on in previous labs are how to design an experiment to answer a question and how to think about the spread in the data you take. To focus on these issues, consider the following task.

> Design a method for launching a marble horizontally from the table onto the floor. The goal is to launch the marble onto a paper target on the floor so that in 10 launches it hits the target **more than 5 times but less than 10**. You may take measurements on the floor or on the table, but may only launch your marble through the air once (before the 10 trials), with TA in attendance.

*Fig.2: An example of a laboratory which focuses on distinguishing between internal and external variation and minimizing external variation.*

## C. Overlooked concepts

There are two measurement concepts that deserve more attention because they are frequently overlooked. The first concept focuses on the difference between a question asking for a prediction of what is likely to happen and a question asking for a description of what already has occurred. The different interpretations can be illustrated by responses to a quiz question that asked the students to compare two sets of data (see Fig. 3). A sophisticated answer to this quiz question might take into account the spread of data, and recognize that the two sets of data overlap almost completely so that the difference in averages is not significant (see Fig. 4).

Which battery lasts longer, Energizer or Duracell?
A student performs an experiment measuring the number of hours two AA batteries from each brand will run a tape player. Her data is below.

| | Trial 1: | Trial 2: | Trial 3: | Trial 4: | Trial 5: | Average: |
|---|---|---|---|---|---|---|
| Duracell (hours) | 11.4 | 12.2 | 7.8 | 5.3 | 10.3 | 9.4 |
| Energizer (hours) | 11.6 | 7.0 | 10.6 | 11.9 | 9.0 | 10.0 |

*Fig.3: The battery lab quiz question.*

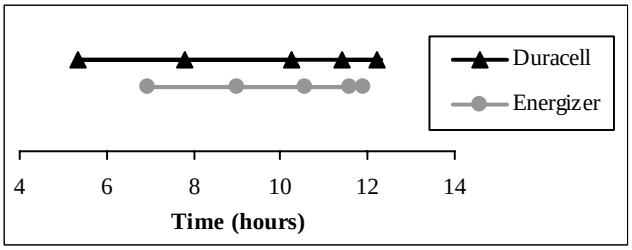


*Fig.4: A representation of the data in the battery lab quiz question. The large overlap indicates that the difference between the two data sets' averages is not significant.*

Consider the following two student responses, where each student counted the number of times each battery lasted longer during a trial.[25]

> Student A: "I would say Energizer because 3 out [of] 5 trials, it lasted for longer hours than Duracell. Also, the average shows it lasting longer by 0.6 hrs."
>
> Student B: "Although on average for the above trials, Energizer appears to last longer, it cannot be concluded that this is the case for most or all situations because in 2/5 trials Duracell was longer and 3/5 trials Energizer was longer."

These two students use the same reasons (Energizer winning 3 out of 5 trials and having a longer average) to reach very different conclusions. We argue that this difference is because they are interpreting the question differently. Student A's response could be interpreted as a description of the data, what happened when the student tested the batteries. Like a sports tournament where best out of five wins, it is true that for those five trials Energizer won. Student B also gives a description of what happened, but takes it further. Student B's statement, "it cannot be concluded that this is the case for most or all situations," could be answering a question asking for a prediction of which battery will last longer in other situations. He appears to claim that one cannot make a prediction; it could go either way. The same question, "Which battery lasts longer?" can be interpreted as a predictive or a descriptive question.

Misinterpreting a predictive question for a descriptive question (or vice versa) can cause instructional difficulties in the student laboratory. Suppose a teacher asks a question, "Does a cylinder with a larger radius roll down the ramp faster?" A student may interpret that question as a descriptive question, and answer "Yes, when we raced them, the larger cylinder won three times, the smaller cylinder only won twice." In an attempt to get students to consider uncertainty, the teacher might respond "But is that difference significant?" This question means nothing to the student – why does significance matter when that is what happened? The teacher has misdiagnosed the student, and the intervention will likely lead nowhere. Instead, the question "if you had an option to bet your life savings on a future race, would you?" may help by introducing gambling, a common situation where people attempt to

predict what will happen, switching the student to a predictive mode.

The second measurement concept can be illustrated by another student answer to the same battery quiz question. This student gave a rather sophisticated response, taking into account the overlap between the two data sets.

> Student C: "The information shows there is significant overlap. So we cannot really tell which battery lasts longer. They last the same approximately."

If the last two sentences of this answer are closely inspected, they may be interpreted as being contradictory. The first, "we cannot really tell which battery lasts longer," could be interpreted to claim that the large range in the battery data is caused by the measurement process. If the range were smaller, perhaps the two data sets would overlap less, and we could see a difference. Possibly when the student tested the batteries, she was not careful to keep everything consistent between trials –she may have played different music at differing volume. Because the researcher's experimental design contributed to the range in the result, it is called *external variation*.

The second statement, "They last the same approximately," could be interpreted to mean that the student was able to take a good measurement and determined that the batteries last about the same length of time. In this case, the batteries themselves cause the range – the student tested some batteries that just happened to last longer than others. This difference may be because the manufacturing process is variable, or perhaps the batteries were stored in different conditions for different periods of time. The source of the range is inside the tested system, so it represents *internal variation*.

To further explain these two ideas, consider the better-known terms *instrumental uncertainty* and *statistical uncertainty*,[26] which also distinguish between sources of uncertainty. Instrumental uncertainty and external variation are effectively synonyms; however, the word instrumental may downplay the role of the person taking the measurement. Statistical uncertainty is by definition random (in contrast to systematic uncertainty) and does not correspond to either internal or external variation. Internal variation, using the battery example, can be random if varying amounts of a certain chemical are placed inside the battery. Internal variation also can be systematic - perhaps the Energizer batteries spent a longer time in transit, reducing each battery's life by the same amount. External variation also can be statistical or systematic.

Internal variation cannot be evaluated in the same manner as external variation. Internal variation is not harmful or helpful; it just exists. If we were to measure the quality at the battery manufacturing plant, we would measure the battery's internal variation while minimizing the external variation. Quality control would then wish to minimize that internal variation. As another example, consider an experiment measuring the energy of gamma rays emitted from a source. There will be a range in the result due to the measuring device (external variation) and the fact that the source does not emit gamma rays with the same exact energy (internal variation). If, instead, we use the gamma source to scatter rays off a target, the differing energy of the gamma rays becomes external variation, and we may include some sort of filtering device to narrow the spread in energy of the gamma rays incident on the target. Whether a variation is internal or external depends on the experimental context.

This difference is critical for questions about quality control or equipment capability. For example, "If you use the device and method described below, how well can you measure a food's weight?" Such questions ask students to reduce external variation while measuring internal variation. Students who conflate the two ideas will be unsure of what to measure, and may treat all variation as "error."

Failing to distinguish predictive questions from descriptive questions and internal variation from external variation is likely to cause learning difficulties for students. In addition, it can lead to miscommunication between instructors and students, and may cause a teacher's intervention to fail. Instructors should be aware of these differences, and laboratory courses should make these concepts explicit. In Sec. IIID we present an activity designed to elicit these and other concepts of measurement.

## D. A concept-based laboratory activity

A projectile lab asks students to build a device to launch a marble off a table and predict how large a target is needed to catch the marble between six and nine times out of ten. The goal is not to build the most accurate launcher, but to determine the accuracy of the launcher (see Fig. 2). The carnival background for the laboratory question provides a context to help students make sense of the question and also authenticates the task of uncertainty propagation.

Students are asked to design an experiment to answer the question, perform the experiment, come to a conclusion, and defend their conclusion to their peers. Most students measure the time it takes the marble to roll across the table, use that number to calculate the velocity of the ball leaving the launcher, and then use that velocity to predict the horizontal distance the ball travels through the air. The shortest and longest times define the largest and smallest horizontal distance. Students then must decide how large to make the target so that it will catch most, but not all of the marbles.

Each stage of the laboratory is critical for students' conceptual understanding. When designing the experiment themselves, students question how much data to take or how to best take a measurement. When performing the experiment, students get a idea of how their measuring instruments or implementation skill affects the uncertainty of their results (for example, how well they can measure time with a stopwatch), and frequently troubleshoot problems such as how to keep the marble rolling straight on the table without experiencing significant friction. When defending their conclusion, students must have an idea of the uncertainty in their data to make a convincing argument for their conclusion.

For the projectile lab example, internal variation is caused by the launching device, which does not always push the marble exactly the same way. External variation is inherent in the various distance (using a meter stick) and time (using a stopwatch) measurements. If students design a reliable launching device, they may predict a larger target than necessary because their range is caused mainly by external variation.

In addition, students are asked to predict the proper target size, not just describe what happened during their measured launches. In some cases, students measure only ten launches, which happen to cluster close together. When they test their target size, the launches happen to differ more, and the target size is too small. Students must decide how many measurements are enough to make a prediction about 10 trials.

This lab is held mid-semester, and requires other measurement concepts. For example, students must decide whether to discard or retain any outlying data, (a concept introduced in an earlier lab) and have to be able to propagate the uncertainty found in their time measurement to the uncertainty in the calculated horizontal distance (a concept newly introduced in this lab).

## IV. RESULTS FROM CONCEPT-BASED INSTRUCTION

In the fall 2002 semester, a computer-based multiple-choice survey was administered to 120 students before and after the described concept-based laboratory course. The survey was based on the Physics Measurement Questionnaire, a free-response survey that tests students' ideas about making multiple measurements in space and time, deciding on a final value to report, and comparing two sets of data.[5] Questions refer to exemplar data "gathered" by releasing a ball from a certain height on an elevated ramp and measuring how far the ball travels horizontally. On the computer survey, students viewed the question, chose an answer, and then were shown different reasons to choose in support of their answer. For the question in Fig. 5, students chose yes or no, and then a pop-up window with the specific reasons appeared. Students could choose multiple reasons, so the percentages shown add up to more than 100%.

The most informative results occurred with the questions asking students to compare two data sets, where the number of students who chose to use range overlap as a reason for their answer increased greatly. For example, in Fig. 5, reasons 4 and 10 use range overlap.[27] Four data-comparison questions were included in the survey, the first two giving students five data points and an average, and the second two giving the average and standard deviation for two data sets. As shown in Fig. 6, there is a significant increase in the number of students using range overlap to compare data sets for all four questions (standard error bars are shown).[28]

Two other groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm.
Their means and the standard deviation of the means for their releases are shown below.

Group A: $d = 434 \pm 5$ mm
Group B: $d = 442 \pm 6$ mm

Do the results of the two groups agree?

**Yes**

1. There isn't a significant difference between the two group's results.
2. Everything has error, it's impossible to get exactly the same every time.
3. There's a difference of eight millimeters between the two group's averages. Eight millimeters is small, and so they pretty much have the same result.
4. Group A's range is from 429 to 439, group B's from 436 to 448, so the ranges overlap.
5. Other

**No**

6. Their averages are different.
7. There is a significant difference between the two group's results.
8. The difference of eight millimeters is a large difference compared to the distances they're measuring.
9. Group A's range has a width of 10 mm, group B's range has a width of 12 mm, so they have different results.
10. Group A's range is from 429 to 439, group B's range is from 436 to 448, they only overlap for 3 mm which is not enough.
11. Group A's average of 434 does not fall within the range for group B (436 to 448), and vice versa.
12. Other

*Fig.5: A standard deviation data comparison question from the multiple-choice computer survey.*

In contrast, consider the results from a traditional, "cookbook" style laboratory associated with a calculus-based introductory physics course. Abbott surveyed 72 students before and after such a course, using a free-response question almost identical to the first data points question we used.[29] For this traditional course (similar to many introductory physics labs), the lab manual gives written instructions on uncertainty estimation and propagation, but rarely required students to perform these calculations. Instead, students frequently calculated the percent difference between two of their results, or between their results and an accepted value. Abbott found no significant difference in the number of students using range to compare data sets after the students completed the course. (One student used range before the course and no students used range after the course.) This result indicates that exposing students to laboratory methods and requiring limited uncertainty calculations will not increase their understanding of uncertainty – it takes a purposeful concept-based laboratory course to do that.

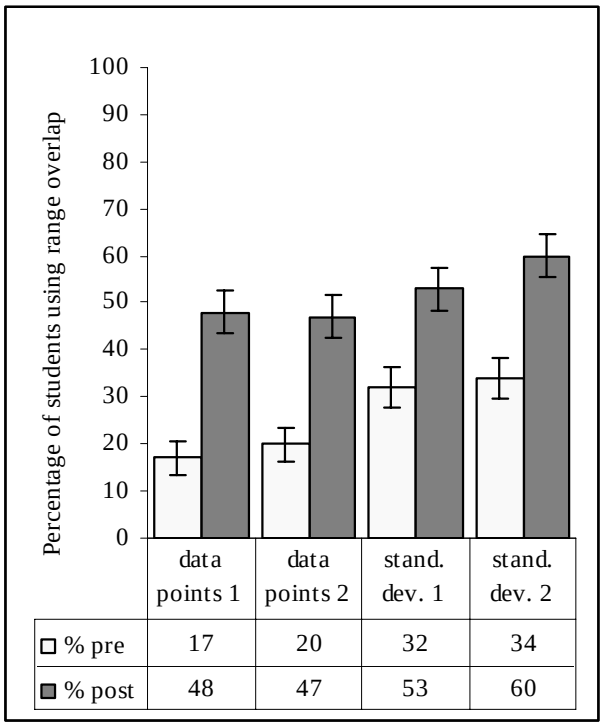


*Fig.6: Percentage of students using range to compare data sets on four survey questions, two giving data points and two giving the average and standard deviation. The standard error of a proportion is shown as error bars.*

## V. SUMMARY

We have argued that students must have a conceptual understanding of measurement and its related uncertainty to plan an experiment, analyze data, and make and defend conclusions. Teaching students to calculate the average, percent difference, and standard deviation often does not lead to such an understanding, and the concepts themselves must be taught explicitly. However, the conceptual instruction of measurement appears to be rare, and instruction on distinguishing predictive from descriptive questions and internal from external variation is almost non-existent. We have provided details of the design and testing of a concept-based laboratory course and have shown how such instruction resulted in gains in students' ability to consider the uncertainty in data when forming conclusions.

## VI. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grant REC-008 7519. I am extremely grateful to Edward F. Redish for his guidance and help, and to Cedric J. Linder for discussions about these ideas. In addition, the professors and graduate teaching assistants involved in the concepts-based laboratory course at the University of Maryland contributed significantly to the course's continual improvement.

[1] M. Welzel, K. Haller, M. Bandiera, D. Hammelev, P. Koumaras, H. Niedderer, A. Paulsen, K. Robinault, and S. von Aufschnaiter "Teachers' objectives for labwork. Research tool and cross country results," Labwork in Science Education, Working Paper 6, Targeted Socio-Economic Research Programme Project PL 95-2005 (1998).

[2] M. G. Sere, R. Journeaux, and C. Larcher, "Learning the statistical analysis of measurement errors," Int. J. Sci. Educ. **15** (4), 427-438 (1993).

[3] J. Leach, R. Millar, J. Ryder, M. G. Sere, D. Hammelev, H. Niedderer, and V. Tselfes, "Students' images of science as they relate to labwork learning," Labwork in Science Education, Working Paper 4, Targeted Socio-Economic Research Programme, Project PL 95-2005 (1998).

[4] F. Lubben and R. Millar, "Children's ideas about the reliability of experimental data," Int. J. Sci. Educ. **18** (8), 955-968 (1996).

[5] 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** (11), 1137-1156 (2001).

[6] P. G. Hewitt, "Conceptual physics - Turning nonscience students on to their everyday environments," Phys. Teach. **10** (9), 522-524 (1972).

[7] For example, see D. R. Sokoloff and R. K. Thornton, *Interactive Lecture Demonstrations* (John Wiley & Sons, New York, 2001), and L. C. McDermott, P. S. Shaffer, and the Physics Education Group, *Tutorials in Introductory Physics* (Prentice Hall, Englewood Cliffs, NJ, 2002).

[8] E. Kim and S.-J. Pak, "Students do not overcome conceptual difficulties after solving 1000 traditional problems," Am. J. Phys. **70** (7), 759-765 (2002).

[9] Consider the popularity of P. G. Hewitt's *Conceptual Physics* (Pearson Education, 2001) as a high school and university textbook.

[10] D. E. Trowbridge and L.C. McDermott, "Investigation of student understanding of the concept of acceleration in one dimension," Am. J. Phys. **49** (3), 242-253 (1981).

[11] R. Scherr, P. S Shaffer, and S. Vokos, "Student understanding of time in special relativity: Simultaneity and reference frames," Am. J. Phys. **69** (7), S24-S35 (2001).

[12] B. S. Ambrose, P. R. L. Heron, S. Vokos, and L. C. McDermott, "Student understanding of light as an electromagnetic wave: Relating the formalism to the physical phenomena," Am. J. Phys. **67** (10), 891-898 (1999).

[13] In this paper we give an example of a course that teaches concepts separately and before calculations. To teach the two simultaneously, see S. Allie, A. Buffler, B. Campbell, F. Lubben, D. Evangelinos, D. Psillos, and O. Valassiades, "Teaching measurement in the introductory physics laboratory," Phys. Teach. **41** (7), 394-401 (2003).

[14] R. C. Hilborn and R. H. Howes, "Why many undergraduate physics programs are good but few are great," Phys. Today **56** (9), 38-44 (2003).

[15] D. L. Deardorff, "Introductory physics students' treatment of measurement uncertainty," unpublished doctoral dissertation, North Carolina State University, 2001.

[16] For similar lists, see Ref. 15, and R. Fairbrother and H. Hackling, "Is this the right answer?," Int. J. Sci. Educ. **19** (8), 887-894 (1997).

[17] This laboratory course, called the *Scientific Community Laboratory*, began in 2001 and continues through Fall, 2004, with a few revisions. We do not claim that this laboratory is the only or the best way to teach the concepts of measurement, but it is an example of successfully improving learning outcomes in this area.

[18] Such reform may need to take the form of a "spiral curriculum." See J. Bruner, *The Process of Education* (Harvard University Press, Cambridge, MA, 1960).

[19] S. Duggan and R. Gott, "What sort of science education do we really need?," Int. J. Sci. Educ. **24** (7), 661-679 (2002).

[20] For example, students consider whether it is better to have several people time the same trial or to run different trials with one person timing.

[21] For a recent discussion on the size of confidence intervals, see D. Groh., "Questioning assumptions," Phys. Teach. **42**, 68-69 (2004); and S. Allie, A. Buffler, B. Campbell, F. Lubben, D. Evangelinos, D. Psillos, and O. Valassiades, "Author's response," Phys. Teach. **42**, 70 (2004).

[22]For example, when measuring whether the period of a pendulum depends on the amplitude of the swing, stacking will not be useful if the amplitude of the swing dies away rapidly.

[23] This issue is difficult. Consider the controversy surrounding what data Millikan chose to publish for his oil drop experiment. See D. Goodstein, "In defense of Robert Andrews Millikan," Engineering & Science, **LXIII** (4), 30-38 (2000).

[24] For example, it is better to start and stop the timer when a pendulum swings through its lowest point (a fast transition) instead of the turn-around point (a slow transition).

[25] The student responses could be interpreted in several different ways. Our analyses may not necessarily be the students' thinking, but we believe that they are reasonable. In either case, these student responses are examples of certain measurement concepts, and need not be interpreted in depth.

[26] For definitions of these terms, see P. R. Bevington and D. K. Robinson, *Data Reduction and Error Analysis* (McGraw-Hill, New York, 2003), 3rd ed., pp. 36-39.

[27] Reasons 9 and 11 in Fig. 5, which also mention the range, do not compare how much the two ranges overlap, and thus are not included as range overlap reasons.

[28] The error bars on Fig. 6 are estimates of sampling error. Assuming a normal population of students studying physics in the U.S., the error bars are an estimate of the sampling errors that could be found if these results were applied to another similar sample of students.

[29] D. S. Abbott, "Assessing student understanding of measurement and uncertainty," unpublished doctoral dissertation, North Carolina State University, 2003.