

# UC San Diego

## UC San Diego Electronic Theses and Dissertations

### Title

Design, Implementation, and Analysis of an Algebra-based Treatment of Measurement Uncertainty

### Permalink

<https://escholarship.org/uc/item/9798w8gn>

### Author

Schanning, Ian

### Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Design, Implementation, and Analysis of an Algebra-based  
Treatment of Measurement Uncertainty

A Dissertation submitted in partial satisfaction of the  
requirements of the degree Doctor of Philosophy

in

Physics

by

Ian Lennon Schanning

Committee in charge:

Professor Adam Burgasser, Chair  
Professor Michael Anderson, Co-Chair  
Professor Stacey Bridges  
Professor Alison Coil  
Professor Douglas Magde

2019

Copyright

Ian Lennon Schanning, 2019

All rights reserved.

The Dissertation of Ian Lennon Schanning is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

---

---

---

---

Co-Chair

---

Chair

University of California San Diego

2019

## TABLE OF CONTENTS

Signature Page: .....	iii
List of Figures .....	viii
List of Tables .....	x
List of Graphs .....	xi
Acknowledgements .....	xiii
Vita.....	xv
Abstract of the Dissertation .....	xvi
Chapter 1 : Introduction & Definitions .....	1
Introduction to Lab Studies: Why should we care?.....	1
Definitions .....	6
1. What is measurement uncertainty? .....	6
2. Range and agreement .....	7
3. Assessment.....	10
4. Lab frame .....	13
Chapter 2 : Course Design .....	17
Introduction .....	17
Course Context .....	17
Uncertainty in the 1-Series .....	19
1. Uncertainty Goals .....	19

2. Defining measurement uncertainty for 1-Series .....	20
3. Using uncertainty to compare results.....	24
4. Treatment of outliers .....	31
5. Propagation of uncertainty .....	33
Structural changes.....	38
1. Learning goals.....	39
2. Pre-lab Questions .....	41
3. Lab Quizzes .....	43
4. Checkpoint questions .....	46
5. Lab conclusion structure .....	49
6. Lab conclusion uncertainty content .....	53
7. TA training.....	60
Chapter 3 : Physics attitudes .....	65
Introduction .....	65
Summer study (full CLASS) .....	69
Evaluating CLASS surveys .....	70
1. Modeling the full CLASS from our survey subset .....	71
2. Contextualizing 1-Series students in Gire’s framework. ....	74
3. Course Context.....	77
4. CLASS individual item results .....	78

Longitudinal study (1AL-1CL) .....	80
Students with previous physics experience.....	81
Attitudes changes within individual 1-Series classes. ....	84
Transition between Winter 1BL and Spring 1CL.....	87
Comparing Spring 1CL to Summer 1CL.....	90
Student workload.....	93
Chapter 4 : Uncertainty probes .....	95
Introduction .....	95
Pilot study on ruler measurement .....	96
1. Results.....	99
2. “Sure Estimators”.....	105
3. Comparing student distributions with expert distributions.....	107
4. TA dependence of the resistor distance. ....	112
Follow-up study.....	115
Why measure multiple times in lab? .....	128
Determining uncertainty from a graph .....	134
Graph uncertainty including propagation .....	138
Comparing two ranges.....	145
Interpretation of uncertainty .....	154
Pixel size: uncertainty from properties of the object. ....	155

Acknowledgements .....	165
Chapter 5 : Responsibilities to biology students.....	166
Biologists in physics labs .....	166
Lab redesigns in physics and chemistry. ....	171
Students perceptions of chemistry and physics labs.....	177
Physics service: a biologist's perspective.....	183
Future work and lessons learned .....	185
1. Center biological content, with consultation with the biology department. ...	185
2. Discuss scientific method / lab practices goals with chemistry / biology.....	187
3. Potential lab course pedagogy changes.....	188
4. Comments on the use of physics attitudes .....	192
Appendix I: Uncertainty and Error Document (Winter 2013 1AL) .....	197
Appendix II: Waves Lab .....	203
Appendix III: Surveys.....	206
Appendix IV: CLASS Item Data .....	222
Appendix V: Data comparison results .....	228
References.....	232



## LIST OF FIGURES

Figure 1-1: Allie, et al. [6] progression of student ideas about experimental data. ....	8
Figure 1-2: Allie, et al. [6] item on rejection of outliers. ....	9
Figure 2-1: "Constant" computer-acquired data. ....	22
Figure 2-2: (Slightly off) triple beam balance weight measurement. ....	23
Figure 2-3: Comparison of ranges (reproduced from Deardorff [8]). ....	25
Figure 2-4: Gaussians of the ranges shown in Figure 2-3 (reproduced from Deardorff [8]). ....	26
Figure 2-5: Comparing a range to a single data point (above) and to another range (below). ....	27
Figure 2-6: Excerpt of uncertainty document showing the how to propagate uncertainties in the case of Ohm's Law. ....	35
Figure 2-7: Learning goals from the first circuits lab in 1BL. ....	40
Figure 2-8: Example pre-lab quiz questions before changes (above) and after changes (below). ....	43
Figure 2-9: Checkpoint questions from 1BL Lab 2 (introduction to circuits). ....	48
Figure 2-10: Schedule of uncertainty concepts from Kung's Scientific Community Labs [5]. ....	54
Figure 2-11: Lab 1 conclusion prompt from 1BL. ....	58
Figure 2-12: Late-course conclusion prompt for 1BL, including the circuit diagram provided in the manual. ....	59
Figure 4-1: Abbott's [12] ruler measurement task. ....	97
Figure 4-2: Pilot study ruler measurement task. ....	98

Figure 4-3: Rubric for student uncertainty techniques. ....	117
Figure 4-4: Rubric for why the class focused on uncertainty. ....	131
Figure 4-5: Determining uncertainty from a graph survey items. ....	135
Figure 4-6: Determining and propagating uncertainty from a graph. ....	139
Figure 4-7: Excerpt from uncertainty document discussing agreement of ranges.....	145
Figure 4-8: Range measurement question and results from Kung's Scientific Community Labs (reproduced from [36]).....	147
Figure 4-9: General rubric used for range comparison questions.....	149
Figure 4-10: Percentage of students making each type of argument. (Totals greater than 100% because some students made multiple arguments.) Pre-instruction (above, N=53) and post-instruction (below, N=208) questions and data shown.....	151
Figure 4-11: Measuring and communicating uncertainty for a pixelated image. ....	156
Figure 5-1: General Biology degree plan at UCSD [41]. ....	169

## LIST OF TABLES

Table 3-1: Sum of squares deviation from prediction for both models. ....	73
Table 3-2: Comparison of expert fraction in CLASS categories between 1CL students and the engineering and first year physics major students studied by Gire [28]. (Std error of 0.05 for 1CL results) .....	76
Table 3-3: Individual CLASS item differences between students who had taken physics before 1-Series and those that did not (pre-instruction). Fractions listed are agreement with expert consensus. Expert consensus is agreement for item 34 and disagreement for the rest. ....	83
Table 3-4: Individual CLASS item differences between students at the end of 1BL and at the beginning of 1CL. Fractions listed are agreement with expert consensus. Expert consensus is agreement for item 34 and disagreement for the rest. ....	87
Table 3-5: Comparison of % expert responses to CLASS items in Spring and Summer 1CL classes. Fractions listed are agreement with expert consensus. P-values in bold are significant using the Holm-Bonferroni test.....	91
Table 4-1: Length estimates and uncertainties for all TAs, expert TAs alone, and pre- instruction students (N=134 for length, N=94 for uncertainty. ....	103
Table 4-2: Uncertainty estimates of the arrow for expert TAs and students. ....	107
Table 4-3: Fraction of students responding that they made multiple measurements in lab to determine uncertainty or spread of data. Significant p-values are bolded.....	130
Table 4-4: Student reporting of voltage. ....	141

## LIST OF GRAPHS

Graph 2-1: Histogram of max-min uncertainty divided by standard deviation for 1000 randomly generated Gaussian data sets of 5 measurements. ....	29
Graph 3-1: Graph of model-predicted CLASS score and the actual CLASS scores, based on Gire’s [28] survey groups. ....	72
Graph 3-2: Gire’s CLASS results reported for each cohort of physics students (from [28]).....	75
Graph 3-3: CLASS 11-question subset results for each class in the sequence. The fraction of expert-like responses is shown; error bars are standard errors of the mean.....	85
Graph 4-1: Histogram of the percentage of students estimating arrow length pre-instruction (N=134).....	102
Graph 4-2: Percentage histogram of pre-instruction (above) and post-instruction (below) student determinations of the uncertainty in the length of the arrow. ....	105
Graph 4-3: Percentage histogram of comparing student and expert ranges (percentages shown), for both pre- and post-instruction surveys. ....	111
Graph 4-4: Comparing K-L resistor difference by lab instructor for post-instruction surveys. (Summer follow-up study taught by the author for comparison) .....	114
Graph 4-5: Percentage histogram (N= 172) for Fall 2012 1AL post-instruction arrow uncertainties. ....	118
Graph 4-6: Self-reported methods for determining uncertainty (N=172).....	119
Graph 4-7: Self-reported methods for determining uncertainty for 0.5 in values (N=62). ....	120

Graph 4-8: Self-reported methods for determining uncertainty for 0.25 in values (N=29)	122
Graph 4-9: Self-reported methods for determining uncertainty for values between 0.02-0.1 in; 0.1-0.25 in, and 0.25-0.5 in, aggregated (N=34)	123
Graph 4-10: Self-reported methods for determining uncertainty for 0.1 in values (N=39).	124
Graph 4-11: Uncertainty values for the minimum increment method (N=41) and the estimation method (N=48).	125
Graph 4-12: Percentage histogram of height estimates (N=219).	158
Graph 4-13: Percentage histogram of uncertainty estimates (N=236).	158
Graph 4-14: Justifications for students who felt their estimate agreed with the provided average and uncertainty estimate. (N=198).	160
Graph 4-15: Justifications for students who felt their estimate disagreed with the provided average and uncertainty estimate. (N=117).	160
Graph 4-16: Percentage histogram of student uncertainty estimates, based on whether they felt their range agreed or disagreed with the provided range of $124 \pm 1$ mm.	163

## ACKNOWLEDGEMENTS

Graduate school has been a long journey. This is a too short list of people and things that helped me get through it.

First, the music: endless repetitions of *Yeezus*, *MBDTF*, *Lemonade*, *The Lion King*, *Persona 5*, *Celeste*, *Hustlin'* and all its remixes, and anything Jay.

Ryan, they can't stop us.

Mike Anderson, I appreciate all that work you put in on top of your own to get me started on this. It was a tough journey at times, but we still put it together.

Nick Koppe. Remember us making those too-jellied sandwiches and we got caught? And in 5<sup>th</sup> grade when they were having a little trouble with my name and you shouted out? I'll never forget that; hope to see you soon.

Joe was there. Always. Thanks so much for everything.

Everybody in Student Affairs: Sharmila, Hilari, Toni and Catherine for helping to keep my day-to-day so pleasant, and for some much general aid and advice.

At QCC: Francesca, Jill, and Kim you're a big reason I feel great here.

Dad, you've provided so much report and encouragement over the years. Thanks for everything and go Lakers.

My whole family, mostly aunts and cousins. Blood-related and not, thanks.

Regina, you made the defense! How crazy is that!

My grandma, rest in peace.

My mom, rest in peace.

Time to go somewhere nice with no mosquitos.

Chapter 4's pilot study is material being prepared for publication as "Implementing algebra-based measurement uncertainty techniques for non-majors". Schanning, Ian L; Anderson, Michael G. The dissertation author was the primary investigator and author of this paper.

Appendix II, in full, is a reprint of the material as it appears in Physics 1CL Lab Manual 5<sup>th</sup> Edition 2014. Schanning, Ian; Salamon, Joe; and others representing the UCSD Physics Department, bluedoor LLC, 2014. The dissertation author was the primary writer of this manual.

## VITA

2006	Bachelor of Arts, University of California, Berkeley
2006-2008	Teaching Assistant, University of California San Diego
2008	Master of Science, University of California San Diego
2009-2011	Research Assistant, University of California San Diego
2011-2014	Lab Teaching Assistant Coordinator, University of California San Diego
2015-2019	Instructor, Queensborough Community College
2019	Doctor of Philosophy, University of California San Diego

## PUBLICATIONS

“Measurement of the Cosmic Microwave Background Polarization Lensing Power Spectrum with the POLARBEAR Experiment”. Phys. Rev. Lett. 113, 021301 – Published 9 July 2014

“Evidence for Gravitational Lensing of the Cosmic Microwave Background Polarization from Cross-Correlation with the Cosmic Infrared Background”. Phys. Rev. Lett. 112, 131302 – Published 2 April 2014

“Assessing student understanding of uncertainty”. Poster, 2015 AAPT Winter Meeting.

Physics 1AL Lab Manual 5<sup>th</sup> Edition 2014. bluedoor LLC, 2014.

Physics 1BL Lab Manual 5<sup>th</sup> Edition 2014. bluedoor LLC, 2014.

Physics 1CL Lab Manual 5<sup>th</sup> Edition 2014. bluedoor LLC, 2014.



## ABSTRACT OF THE DISSERTATION

Design, Implementation, and Analysis of an Algebra-based  
Treatment of Measurement Uncertainty

by

Ian Lennon Schanning

Doctor of Philosophy in Physics

University of California San Diego, 2019

Professor Adam Burgasser, Chair

Professor Michael Anderson, Co-Chair

Learning to measure and propagate uncertainty is a necessity for understanding experimental data. But because statistics and calculus are required for a rigorous analysis of uncertainty, students in algebra-based physics classes are rarely able to determine the meaning of their experimental results. This thesis describes the design and implementation of an algebra-based

method for calculating, propagating, and determining agreement with measurement uncertainty that was taught at scale to students of introductory physics for biologists. Changes to the previous lab implementation are described, including the additional of learning goals, in-class checkpoint questions, lab write-up prompts designed around uncertainty, and TA training. Student attitudes towards physics are measured in different instructional circumstances using the Colorado Learning About Science Survey. Several different methods of measuring how well the students learned the technique and the meaning of uncertainty are analyzed. Students showed more comfort and ability in using estimation and spread to determine ruler lengths after instruction. Students were more likely to use spread-based approaches in computer collected data. Students who performed the uncertainty propagation procedure were more likely (97% to 28%) use their calculated uncertainty in data comparisons. Different framings of uncertainty are compared. Since biology students also take lower division lab classes in chemistry, the complementary approach that each science takes in lab instruction is explored from a course design perspective and via student interviews.

## **Chapter 1 : Introduction & Definitions**

### **Introduction to Lab Studies: Why should we care?**

Laboratory classes have been a part of the introductory physics experience for more than fifty years, even as different instructional emphases have come and gone (see Trumper [1] for a general overview). Nevertheless, Arons [2] found that the goals for these classes fell into a few broad categories:

- a) To verify or confirm laws, relations or regularities asserted in text, class, or lecture
- b) To have some experience with actual physical phenomena
- c) To have the experience of, and develop some skill in, handling instruments and making significant measurements
- d) To have the experience of planning and doing experiments and thus encountering some of the “processes of science”
- e) To learn something about minimizing error and about the treatment and interpretation of experimental data

Notably, only the first category (or, perhaps the second in the case of exceptional in-class demonstrations) is common in lecture classes, requiring the laboratory classes to do the heavy lifting associated with the later categories. Typically, physics lecture classes are filled to the brim with physics concepts, mathematical models, and word problems including real-world context. Even though these ideas may have taken dozens or hundreds of years to develop and refine, students are expected to learn how the concepts work and how to work the models in a handful of weeks. Learning physics content is learning the established models created long ago.

At the same time, these later objectives are broadly applicable to any science education, as they represent the actual practice of science, rather than the ideas that previous scientific inquiries have generated. Prospective scientists, in other words, at some point must eventually *do* science as opposed to simply reading and comprehending the results of previous scientists.

Though science degree programs encourage undergraduate research opportunities (and upper division lab classes) whose goal is to introduce scientific methodology in a more rigorous way, students should be exposed to scientific epistemology even in their initial coursework. Before undergraduate students have the opportunity to be mentored by a practicing scientist, they are students in introductory classes that can help the process along by laying the groundwork for science practice.

Unfortunately, it is rare for these practices to be encouraged in the lecture portion of a physics class, especially introductory class. Most textbooks have a token few pages on measuring and uncertainty that is easily ignored and often skipped; moreover, the established values for physical constants are quoted without any uncertainty. This encourages the idea that science is handed down from an authority, counter to both scientific and educational best practices.

These issues can contribute to the feeling among non-experts that, as the Colorado Learning Attitudes about Science Survey (CLASS) puts it, “The subject of physics has little real-world relation to what I experience in the real world.” There is a danger in presenting the physical results as exact (i.e., lacking measurement uncertainty), while simultaneously being counterfactual to what students experience in the real world. For example, according to the textbook Serway and Jewett [3], “It is well known that all objects, when dropped, fall toward the Earth with nearly constant acceleration”. Outside of the classroom, relying on their own life experience, students

know that some objects (a hammer) will fall to the ground faster than others (a feather). While a good textbook may note this fact, it does not replace the experience of actually constructing an experiment to measure  $g$ , especially in the context of introductory physics lab materials and methods. Here, the variability and measurement uncertainty is unavoidable, and the previously abstract effects of friction drastically alter the final result. By contrast, in the rare occasions that Serway does mention uncertainty, it is for very precise experiments, where an uncertainty of  $\pm 0.3\%$  is “relatively large”; this is unlikely to be a large uncertainty for any lab experiment performed by an introductory student. Rather than presenting the production of scientific data as something mysterious with seemingly impractical precision, a well-designed lab can give students a chance to produce some science of their own.

Of course, the practicing scientists who often teach these courses know this. This helps to explain why laboratory classes, despite logistical and cost issues, remain a core part of introductory physics instruction. A simple view is that labs serve as the “practice” that complements the “theory” of the traditional lecture.

In introductory physics instruction, then, the topics of measurement and uncertainty (which remain central to scientific inquiry) are relegated to laboratory instruction, especially in the case of classes for non-physics majors. If the lab curriculum of these students does not focus on these topics, they will not be present in their physics instruction at all<sup>1</sup>.

Of course, lab classes don’t necessarily support those scientific literacy goals by default either. For example, “cookbook labs” are characterized by: spelt-out procedure, repetitive data-

---

<sup>1</sup> Moreover, since chemistry and physics are often prerequisites for biology, these students may not encounter any real scientific inquiry at all until their upper division biology lab classes! See Chapter 5 for further discussion on the role of chemistry and physics lab classes for biology students.

taking, and an overall focus on process rather than sense-making. Cookbook labs are thus analogous to traditional lecture in the usual physics education research (PER) literature, as a barely functional but common class option from which one should deviate to support a good learning environment. The detailed written procedure of these labs cannot model the generation of scientific questions, which of course are not given to the scientists in a lab book; further, it cannot model the connection of those questions to the experimental design which remains a core part of scientific practice [4]. Moreover, the cliched summary question of such labs – “do the results agree with what you expect?” – can only contribute to sense-making if students can answer that question in a quantitative and scientific way.

What, then, is the laboratory analogue to the “flipped class”? What would a lab course designed from scratch to focus on how scientists reason and measure look like? One example of such a vision can be found in the Scientific Community Labs built by Kung (nee Lippmann) [5]. Not only did the students in the class design experiments, perform measurements and analyze their results, but they also communicated the whole process to the rest of the class. Kung’s goal for the class was that it could serve as a scientific community in itself, modeling in detail what real physicists do to model, measure, and communicate their work. See Chapter 2 for more details on this extraordinary course.

That said, Kung’s laboratory class structure cannot be easily emulated by most instructors teaching introductory physics. Her labs represent not only a great deal of personal effort (implementing and measuring them was her PhD dissertation), but also significant administrative and departmental buy-in from the University of Maryland and her advisor, PER luminary Edward Redish. By contrast, instructors at many colleges and universities may find that even the text of their lab manuals are fixed, or at least subject to departmental approval; this limits the scope of an

ambitious researcher. At some universities (like UCSD), lecture and laboratory sections are mixed for a given cohort of students, meaning even a small scale experiment that combines lecture and lab would be difficult or impossible to organize. Finally, a graduate thesis represents the focused work of several years, while an instructor at a community college might have a semester of sabbatical to design and organize changes to laboratory curriculum.

**The goal of this thesis is to apply what has been learned from studies like Kung's to a more common teaching setting.** At UCSD, we were constrained by the pre-existing cookbook-like lab manuals; the separation of lecture from lab; the number of students (often around 900), which required a small army of teaching assistants to teach lab classes each quarter; and, in the case of the 1-Series, teaching biology majors a full year of physics. This thesis serves as a case study on how inquiry elements (with a special focus on measurement uncertainty) were implemented in non-major lab classes still reliant on cookbook-like lab manuals. In particular, it describes the implementation of an algebra-based system for determining and propagating uncertainty, and a way to use uncertainty ranges to compare results without the need for more advanced statistical training that many students lack.

The rest of this chapter provides an introduction to the topic of measurement and uncertainty in the lab-focused PER literature. Chapter 2 describes the context of the study, including the student population, course content, and specific implementation of the new pedagogy. Chapter 3 describes the effect of lab instruction on the students' attitudes towards physics, as measured by the CLASS. In Chapter 4, various student measurement and uncertainty skills are assessed. Finally, Chapter 5 discusses the complementary role that physics and chemistry lab classes have in serving biology students, as well as plans for future investigations.

## Definitions

Considering the laboratory's role as an “experimental” complement to the “theoretical” lecture, it is no surprise that there is a focus in the PER literature on the process of measuring, including the interpretation and sense-making of results. In this section, the literature for three aspects of data interpretation is reviewed, including: measurement uncertainty; its use in comparing data results; and how students' lab experiences filter and process the previous two.

### 1. What is measurement uncertainty?

Much of the trouble in experimental physics is not only measuring a best estimate, but also quantifying how “good” this estimate is. This includes both statistical uncertainty (how is your best estimate affected by random chance during the measuring process?) and systematic uncertainty (how do your measuring equipment and process bias your results?). Generally, determining both uncertainties for a given experiment is a significant endeavor, including advanced statistical analysis, modeling of equipment, and simulations. Uncertainty analysis is crucial enough that a common shorthand of the validity of an experiment is not its effect size but how many standard deviations it is away from the null result.

For our purposes, measurement uncertainty (or after this point, just “uncertainty”) is a quantity attached to the best estimate that communicates the spread or range of that estimate. It can serve as an indication of the quality of the design or execution of an experiment, but is mostly used to compare the results of one experiment to another. Using the uncertainty, one can calculate



how likely it is that the best estimates from each experiment are truly different, as opposed to just being measurements of the same underlying quantity.

Introductory lab students also run experiments in their class time; their results also need some kind of uncertainty analysis to be meaningful. Although lab students' grasp on the machinery of uncertainty and analysis will be less complex than that of actual working scientists, this still leaves open many options for instruction: should students measure uncertainties at all, or should they rely on significant figures provided by the instructor or measuring device? Should students use detailed statistical tools, like standard deviations and t-tests? Should they only calculate statistical uncertainty of their results, or attempt to quantify systematics in their measuring devices? Before explaining the approach that was chosen for this study in Chapter 2, it is important to examine how students make use of measurement uncertainty, as studied by the literature.

## **2. Range and agreement**

Unlike typical introductory physics problems where the key quantities are provided in the problem description or on an equation sheet, laboratory-based physics work requires the direct measurement of the environment before calculations and comparisons can occur. There is a fundamental difference between the given quantities in the textbook and the measured quantities in a laboratory; while the former are exact, yielding precisely reproducible results that make it easier to grade problems, the latter are inherently variable because of measurement errors and systematics. In real life, remeasurements will produce different results; rereading a textbook problem, by contrast, will never change the numbers therein. Transitioning from textbook physics

to laboratory physics requires students to understand that this difference between assumed, exact quantities and real-world, variable ones.

One might expect, then, a continuum between treating measured quantities as exact and as an ensemble of many different measurements. Allie, et al. [6] provided an extension of the framework of Lubben and Millar [7] which presented student understanding on an ascending scale from a fully point paradigm (where students measure once and accept that as the correct answer) to a full set paradigm that encompasses several different types of procedural knowledge: taking multiple measurements to establish a mean & spread; using the spread to determine the quality of the result; rejecting outliers; and using spreads to determine the consistency of multiple sets of measurements. Even though this is a strong system for classifying student responses to questions about uncertainty, it is less effective when used as a classification of student ideas when they are following lab procedures in a classroom.

**Table 1. Model of progression of ideas concerning experimental data.**

<i>Level</i>	<i>Student's view of the process of measuring</i>
A	Measure once and this is the right value
B	Unless you get a value different from what you expect, a measurement is correct
C	Make a few trial measurements for practice, then take the measurement you want
D	Repeat measurements till you get a recurring value. This is the correct measurement
E	You need to take a mean of different measurements. Slightly vary the conditions to avoid getting the same results
F	Take a mean of several measurements to take care of variation due to imprecise measuring. Quality of the result can be judged only by authority source
G	Take a mean of several measurements. The spread of all the measurements indicates the quality of the result
H	The consistency of the set of measurements can be judged and anomalous measurements need to be rejected before taking a mean

*Source:* Adapted from Lubben and Millar (1996).

**Figure 1-1: Allie, et al. [6] progression of student ideas about experimental data.**

The detailed lab procedures from “cookbook” or “cookbook with inquiry” labs already mandate the taking of multiple measurements, since they are written by scientists who understand that a single measurement is not enough to establish a spread. In itself, this cuts off A & B from the classification scheme, since students will always take multiple measurements because they are following the instructions. In my experience, students will automatically discard (what they perceive as) outliers when taking data, to the point that I felt the need to emphasize that students should keep all data they take, at least initially. For students that don’t grasp why measuring a spread of values is indeed the goal, there is a tension between following the instructions (and recording multiple different values) and following your instincts to keep the repeated or best or central values. They can think of the spread as evidence of mistakes (avoidable or not) in measuring (*errors* rather than *error*) that may betray some procedural failing that might need to be fixed. While this can be true in some cases, I see the baseline role of the TA in laboratory class as preventing those sorts of gross procedural errors that would drastically warp the results, while leaving in reasonably-sized procedural missteps that help to establish a realistic spread.

*Perceptions on spread and comparison of data sets*

The first probe on spread of measurements dealt with how to handle an anomaly (AN):

A group of students have to calculate the average of their (distance) measurements after taking six readings. Their results are as follows (mm): 443, 422, 436, 588, 437, 429.

The students discuss what to write down for the average of the readings.

**A:** ‘All we need to do is to add all our measurements and then divide by 6.’

**B:** ‘No. We should ignore 588 mm, then add the rest and divide by 5.’

**Figure 1-2: Allie, et al. [6] item on rejection of outliers.**

As an example, consider Allie's questions about rejecting outliers, used to judge whether students had reached step H in their hierarchy, as shown in Figure 1-2. To simplify the distinction, Allie listed an outlier that was beyond the pale from a statistical perspective: the 588 mm reading is nineteen standard deviations away from the mean of the other five values. (Practically speaking, this sort of result rarely happens in the context of lab class data analysis, as students would typically drop the value themselves without even recording it). While most (56%) of the students believed the point should be excluded, 42% believed the point should be kept. Those who kept the point did so either for procedural reasons (need to include all data when taking the average) or because the outlier value represented the spread that the group actually measured and should be included.

The goal of physics 1-series labs is not to collect the best, most accurate and precise data, but to encourage proper analysis of (flawed) data; consequently, analysis of included outliers is preferable to their nonexistence. More discussion on how uncertainty was measured and used in our labs can be found in Chapter 2, along with the details of how this pedagogy was implemented.

### **3. Assessment**

Once the uncertainty skill sets have been identified, it is important to craft a way to assess those skills to determine how successfully students have learned them. Deardorff [8] was the first to attempt a comprehensive study of student understanding and skills for measurements. He developed included a laboratory practicum to test both the students' measuring ability and their use of the resulting uncertainties. For instance, students directly measured the diameter of a penny with physical tools. By providing different options for measurement tools in the physical test (i.e.,

calipers vs a normal ruler), Deardorff was also able to test the confidence students felt with the different tools.

Unusually, Deardorff's thesis applied these methods to an educationally diverse set of students: algebra-based biology majors, engineering students and introductory physics majors all completed the same surveys and had the opportunity to show the same practical measurement skills.

While some laboratory skills apply regardless of the topic of the lab class (working in a group, properly calculating and reporting measurement uncertainty), others depend more on the majors and career goals of the student population. For example, an introductory lab class that feeds an electrical engineering program might choose to spend a few weeks studying op-amps, while that would be an inappropriate choice for biologists. The practical laboratory skills necessary for biology would be better assessed in a biology or perhaps a chemistry lab class.

Allie, et al.'s [6] instrument for measuring student understanding of uncertainty was the Physics Measurement Questionnaire, a series of probes like than in Figure 1-2 which measure the role of uncertainty in data collection, analysis and in the comparison of data. Elements of the PMQ in whole or modified have been used to assess these three major areas of the use of uncertainty in [cite here in order], for example.

Pollard et al. [9] applied the PMQ to a traditional introductory lab course, which provides a model for where student reasoning about uncertainty lies in such courses how lab instruction affects student ideas. Encouragingly, student performance on the PMQ overall became significantly more set-like after instruction, with 34% of students reasoning consistently with set-like reasoning before instruction rising to 56%. Most of this positive change was due to improved

set-like reasoning on the probes [10] related to deciding if two ranges with different means agreed. The fraction of the students using spread reasoning on those probes rose significantly from 45% to 60%. In the probe studying the connection between making multiple measurements to uncertainty, however, although fewer students used point reasoning after instruction, there was no corresponding increase in consistent set reasoners, which stayed around 35%. The authors conclude that set reasoning in the area of data collection is a useful place to focus for pedagogical changes in lab, since there is significant room for improvement.

Majiet and Allie [11] further explored data collection reasoning by expanding from the PMQ to try to determine why conceptual shifts in the use of uncertainty occurred during instruction. In developing their probe, they classified student responses according to how well explained student written responses were to their probe items. Students were less able to elaborate on the probe relating to how repeated measurements were related to an uncertainty or spread.

Besides calculating uncertainty from a range, Deardorff's dissertation [8] also addressed students estimations of uncertainty, which a poll of professors considered a key skill. While students tended to estimate the uncertainty in a reasonable way for direct measurement (e.g., measuring the length of an object using a ruler), they were much less comfortable with uncertainty estimates from calculated quantities. For example, they often quoted uncertainties to three or four significant figures (despite using significant figures correctly for directly measured quantities); moreover, the uncertainty estimates themselves were less reasonable. This points to the need for a propagation procedure, to help ensure that uncertainties for these derived quantities are accurate. Abbott [12] followed up on this study by developing an survey tool for ruler estimation of uncertainty that is discussed in more detail in Chapter 4.

Day et al. [13] further explores the use of uncertainty in calculations via a context-rich problem where students were asked to derive the weighted average when given data comprising the size of ostrich eggs along with the uncertainties in those measurements. After students came up with their own methods, the reasoning for the proper weighting (i.e., weighting the measurements with their inverse uncertainties, and properly normalizing the results) was explained to them. Later, transfer of this approach was measured in comparing two different student measurements required for a calculation, where one had significantly more uncertainty than the other. Students were significantly less likely to ignore the uncertainties entirely in the second activity (30% to 13%) if they worked on the ostrich egg measurement problem that semester. They conclude that even if students are not able to remember the detailed calculation method using the weighted average later on, they are still more likely to recognize that examining the uncertainty of data sets (rather than just their average values) is an important consideration.

The literature on the use of uncertainty in the lab covers a wide variety of skills. Consequently, the procedural assessment tools that were developed for study address each of the following: how to calculate and estimate uncertainty; how to use that uncertainty to compare results; and how to propagate that uncertainty. The details of these assessments can be found in Chapter 4; the surveys themselves can be found in Appendix III.

#### **4. Lab frame**

Even if students can calculate a measurement uncertainty and use that uncertainty to compare results, the ultimate goal is for them to *do science*, which is as much a mindset and approach as a procedure. Beyond skill building, students should to understand the “whys” of the

skills that they are using. A student that simply perfectly memorizes and recites every concept stated in a classroom is much less a scientist than one who carefully reflects on and questions each, even if the latter makes more mistakes.

Context can help to dictate the frame of mind a person uses to interpret a situation. For example, a velocity of 3 m/s written in a textbook problem means something different than the same velocity as measured on an air track. The former is exact and does not need further examination before it can be used in a calculation, but the latter requires several more measurements before it can be properly interpreted. Still, there are many possible lab frames that can be encouraged or discouraged by the provided lab material. Before the changes in lab were made for this study, the use of uncertainty in the class was limited to a short “error” section at the end of each lab write-up where students speculated on possible things that could have gone wrong in the experiment. Even if students had a more advanced understanding of measurement uncertainty from previous classes, that sort of framing hardly primes them to use it.

One of the more common failure states is a “following directions” lab frame, where students simply follow the directions in the lab manual without reflecting on their data. Kung, in her study of lab metacognition [14], termed this a logistical mode of operation, as students were focusing more on the step-by-step operations of making the lab work than making sense of their data. Kung examined three different styles of lab (cookbook; cookbook with explanations; and her own scientific community labs, an inquiry-based approach discussed further in Chapters 2 and 4), tracking how often students spent time discussing logistics compared to sense-making. Of course, in labs where students are responsible for designing and implementing their own experiments (rather than following set directions), she found that they spent more time in a sense-making mode. Although students working with cookbook labs tended to narrowly focus on logistics, simply



asking “why?” type questions in the manual could sometimes prompt them to shift to the sense-making mode.

It is especially crucial for students to be in a sense-making frame as they make use of measurement uncertainty to compare data to expectation. Viewed outside that frame, calculating uncertainty is just another mechanical task that students follow because they are told to, just as they might try to implement any arbitrary instruction listed in a lab manual. This informed our design of mid-lab checkpoint questions and post-lab conclusion questions; examples of each of these can be found in Chapter 2.

Holmes and Bonn [15] used interviews as a way of assessing student lab frame in the context of comparing three different types of measurement of the index of refraction (using Snell’s Law, Brewster’s Angle, and Total Internal Reflection), where one of the measurements had a large systematic error. They found that certain ways of comparing results commonly used in science that could help catch the systematic error (e.g., comparing results with other groups, or with a reference value) were viewed as cheating by the students. Moreover, they found that although students successfully performed the calculation procedures involving uncertainty, they were less likely to reflect on what the results of those differences actually meant.

In this dissertation, the lab frame of students was also assessed through the use of interviews, as will be seen in Chapter 5. Biology students at UCSD tend to take both chemistry and physics lab classes before laboratories in their major, so the default lab frame of students is initially set by these out-of-major lab classes. Interestingly, because the physics and biology labs have quite different goals and structures, the optimal lab frame for each class differs as well. Chemistry lab classes, for example, focus much more on the logistical mode since they are

designed to teach biology students specific experimental skills that they might find useful in subsequent biological research. Indeed, the students consciously recognized the difference in lab frame encouraged by both lab classes. I argue in Chapter 5 that these two introductory approaches complement each other, as chemistry and physics lab instruction together provide a more complete scientific experience for the biology students than either would alone.

## **Chapter 2 : Course Design**

### **Introduction**

Before describing the results of this study, it is necessary to detail the previous context of the physics lab courses at UCSD, and how and why those classes were changed. While many instructors would like to improve the way that uncertainty and error is handled in their lab classes, each of their institutions has different lab documents, flexibility, support and bureaucracy affecting the speed and depth of those kinds of changes. By fully detailing the goals, constraints, and execution of the new pedagogy, I hope to provide a clear picture of why each change was made for this study to better allow others to apply our lessons to their own, personal context.

First, this chapter describes the general context of the physics courses, including the student population. Next, it examines the approach taken towards uncertainty, including comparisons of results and propagation. Finally, the new design of the laboratory courses is explained including how the pre-existing conditions at UCSD constrained those choices.

### **Course Context**

At UCSD, introductory physics instruction is divided into four different tiers: a one course conceptual physics class; a three-course series for biology majors (1-Series); a four-course-series

for engineers and other physical scientists (2-Series); and a five course-series for physics majors (4-Series). As opposed to the more calculus-heavy 2-Series and 4-Series, the 1-series lab courses are mostly algebra-based, with an occasional light touch of calculus mostly for in-class derivations. These courses were composed almost entirely of current or prospective biology majors and were the object of all of the pedagogical changes of this study.

UCSD is a large, nationally ranked university that is especially well-known for biology. Undergraduate enrollment in the first year of this study was over 22,000; biology majors represented around 20% of declared majors at UCSD over the last decade [16]. The school is currently ranked eighth worldwide by U.S. News and World Report for biology [17]. Although the university primarily serves California, around 20% of its undergraduate population were out-of-state or international students [17].

Each 1-Series lab course was a co-requisite with the corresponding lecture course (i.e., Physics 1AL was a co-requisite for the lecture class Physics 1A). A single instructor oversaw all three lab courses, whereas each lecture course was divided into multiple sections, each taught by a different instructor. Typically, there was little coordination between the multiple sections of a given lecture course, or between the lectures and their corresponding lab class.

Because each biology major was required to complete the entire year-long physics sequence, the total enrollment of the combined labs in a given quarter was around 900. These were split up among the three courses (1AL, 1BL and 1CL), and further subdivided into 26-student lab sections, each overseen by a teaching assistant (TA) or two, depending on the quarter. Teaching assistants were primarily physics graduate students, often in their first year. In the fall where enrollment was highest, there could be as many as twenty graduate students teaching 1-Series lab

classes; from their sheer volume, these labs often served as the initial teaching assignment for first semester graduate students.

Despite the enormous student total, the entire 1-series lab system was overseen by a single instructor, who was ultimately responsible for the final grades of each of the three lab courses. Consequently, it was difficult for the instructor to individually organize the teaching assistants for each class. Each lab course was overseen by one or more Lab Teaching Assistant Coordinators (LTAC), whose primary job was to coordinate and train the teaching assistants that would teach the individual lab sections. LTACs were drawn from more experienced graduate students that had previously taught the labs. They were also responsible for overseeing a few office hours a week to help students working on lab assignments.

Student work was graded by the TA who taught their lab section. Students were assessed based on a combination of pre-lab assignments, reading quizzes, post-lab writeups, and attendance. Written work was graded without any unifying rubric; while attendance and similar policies (no make-up labs, etc.) were determined by the overarching (faculty) lab instructor.

## **Uncertainty in the 1-Series**

### **1. Uncertainty Goals**

The inspiration to change the lab classes at all arose from observations of student lab write-ups. Uncertainty was limited to a sentence or two about “error” at the very end of the document, mostly consisting of students guessing why the experiment failed, often because of “human error”.

Worse, the data section of the write-up was similarly vague: students simply decided whether the results looked close enough to the intended value. This was not surprising: there was no instruction devoted to how to measure or estimate experimental uncertainty, and then to use those values to figure out what the data was saying. Consequently, data comparisons had to be made using student intuition rather than any quantitative approach.

We decided that after taking the lab course, students should be able to do the following:

- Calculate or estimate the measurement uncertainty for *each* measured quantity
- Interpret this uncertainty as a measure of variability of the quantity
- Judge the consistency of multiple data sets (or a single data set with a theoretical quantity) using this variability
- Propagate uncertainties to facilitate meaningful comparisons, if necessary

These goals, written as they are, could apply to variety of different courses at different levels of mathematical sophistication. In the 2-Series (engineering), for example, students work towards similar goals by calculating standard deviations, using basic statistical tests like t-tests, and propagating uncertainties using partial derivatives. Because 1-Series labs were algebra based, a different system had to be employed; while the math might have been barely appropriate for that level, there was no recitation time to be able to teach a system of that complexity. Instead, a more intuitive system was developed that allowed students to estimate uncertainties and use them in the propagation of errors without doing mathematically challenging partial derivatives, or even ponderous (or black-box) calculations of standard deviations.

## **2. Defining measurement uncertainty for 1-Series**

Commonly, 1-Series lab procedures instruct students to measure a quantity multiple times: usually around 3-5. Using those data in effect as a single measurement of the quantity, those values

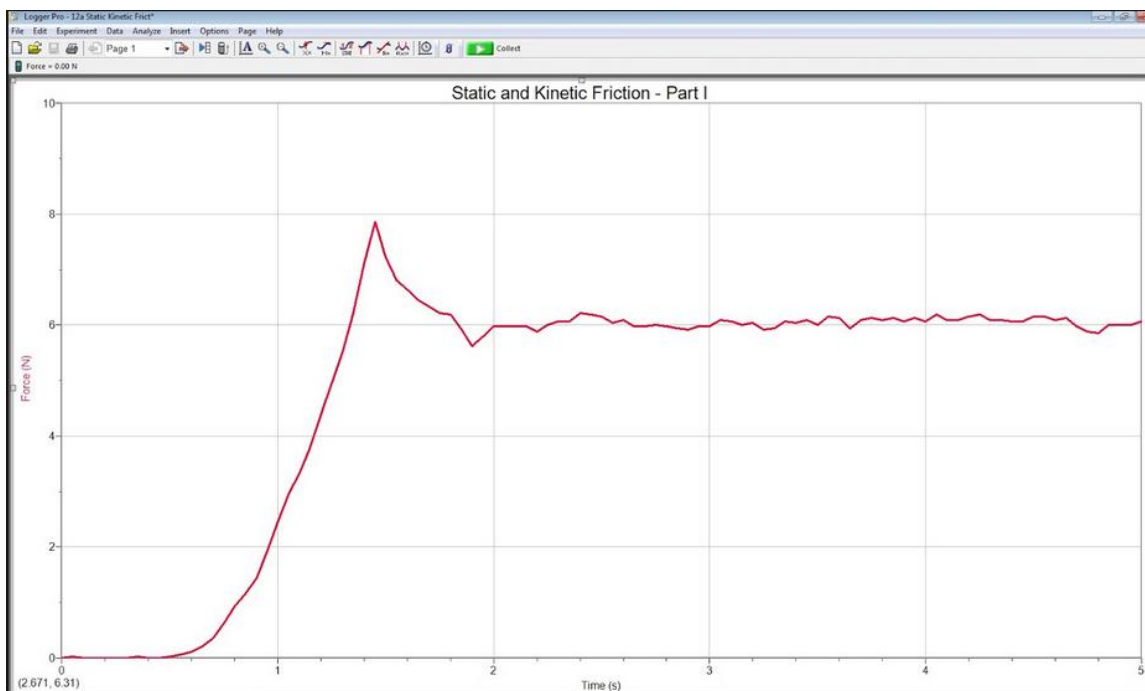
define the range of measured values for that quantity; the largest value is the maximum of the range, and the smallest value is the minimum of the range.

The *uncertainty* of the measurement is defined as the difference between the maximum edge of the range and the minimum edge of the range, divided by 2. For example, for the measured values [3, 9, 5, 7], the range would be from 3-9, and the uncertainty would be  $\pm 3$ . The full set of measurements can be defined either by the range, or by the mean  $\pm$  the uncertainty. For the above example, these methods yield: 3-9 or  $6 \pm 3$ . Alternately, the uncertainty can be calculated by first calculating the mean, and then subtracting either of the two edges of the range from it; uncertainty is always defined as a positive number<sup>2</sup>. This uncertainty represents how repeatable a measurement of that quantity is under the current conditions.

In some cases, data measurements were recorded by computer based data acquisition (DAQ) systems like LoggerPro, which typically records many more than five measurements per quantity. Here, too, students can graphically record the highest and lowest value in the appropriate period of time, and use those upper and lower limits to calculate the measurement uncertainty. Even though the data in Figure 2-1 on the right hand side appears to be basically constant overall, a value of the measurement uncertainty can be estimated as  $\pm 0.2$  m/s.

---

<sup>2</sup> This procedure for calculating uncertainty will lead to a slightly different range than simply using the largest and smallest values if the data is not symmetric. These differences are always smaller than the uncertainty itself, and so our method does not distinguish between these results. As will be seen in data comparison section of Chapter 4, the extreme edges of a range are merely suggestive of the true possible boundaries of the measured quantity, especially in typical lab conditions of 3-5 measurements defining that range.



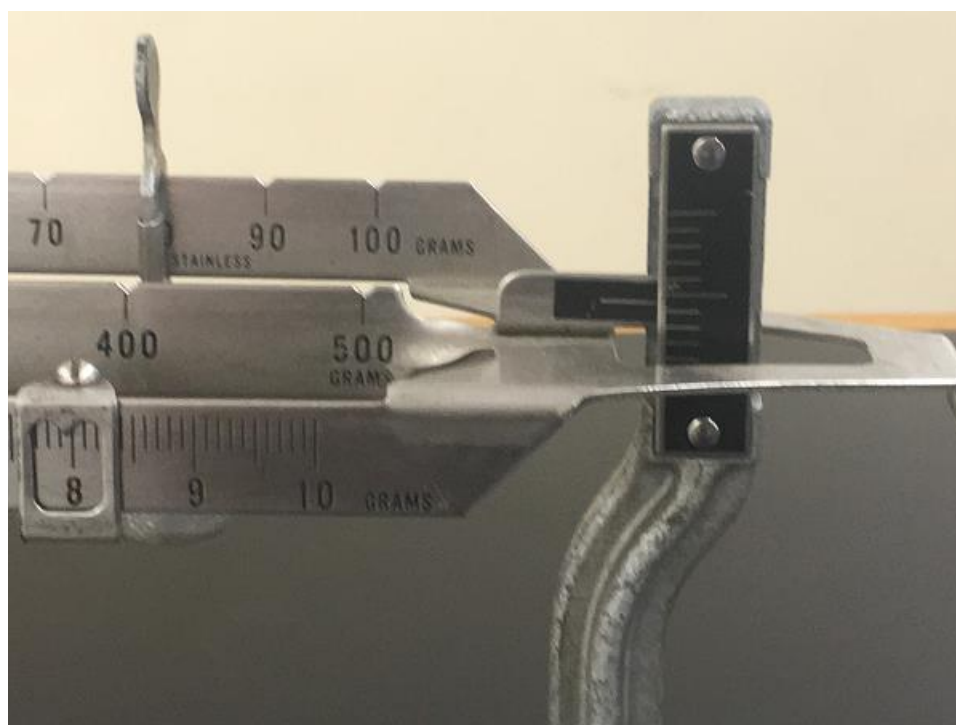
**Figure 2-1: "Constant" computer-acquired data.**

A range can be defined anytime that multiple measurements of a quantity are made with identical conditions. Sometimes, however, a quantity can only be measured once. In this case, students estimate the uncertainty of the measurement directly, and then use that uncertainty to define the likely range for the quantity. In the case of digital measuring devices, the fluctuations of the measuring implement also serve to define a minimum and maximum of possible values, allowing students to record a range and calculate an uncertainty.

In some cases, such as with a triple beam balance, students can estimate uncertainties by physically playing with the measurement device to determine its responsiveness to changing conditions. For the measurement uncertainty of a correctly-zeroed balance, the key insight is how small of a margin between the measuring lines is "close enough" to call the weights in balance.. No matter the technical precision of the balance, an experimenter is eyeballing the lines to see



when they coincide; since the lines indicating balance don't ever exactly line up, the experimental uncertainty depends on how much variation is allowed before trying to balance the scale again. To those with a careful eye, even a miniscule misalignment of the counterweight line might count as “not balanced”; for less precise measurers, the scale might be considered “basically” in balance for up to 0.2 or 0.3 g. Regardless of the numerical value, it is important that students can record the appropriate size of the uncertainty for their experiment.



**Figure 2-2: (Slightly off) triple beam balance weight measurement.**

This idea of measurement uncertainty is necessarily less specific and mathematically detailed than the usual idea of a standard deviation, calculated by comparing each individual measured value to their mean. See Taylor [18] for an example of this approach for data analysis at

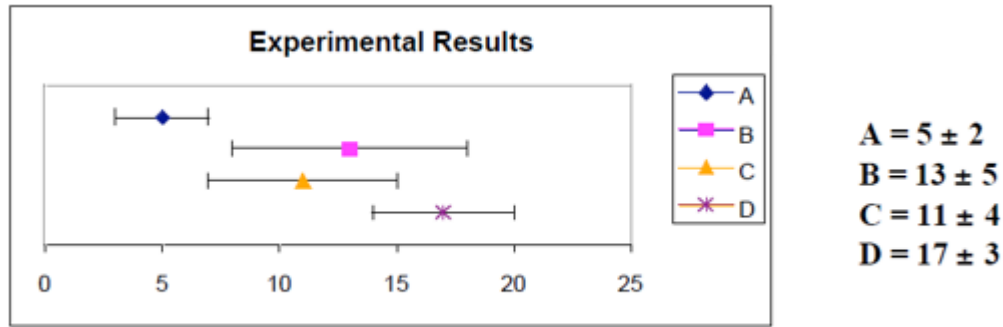
an introductory level. The standard deviation was not chosen as the measure of uncertainty for the following reasons:

1. The calculations are laborious to calculate by hand. Alternately, they require the use of Excel or other similar programs. In the latter case, the process for determining the standard deviation becomes a black box.
2. The explanatory power of the standard deviation is limited in a lab context where at most five measurements are made for a given quantity.
3. The standard deviation is a means, not an end. Using the standard deviation properly requires a further suite of statistical tests, with significant time spent both to explicating these tests and applying them in the appropriate context.

### **3. Using uncertainty to compare results**

An overlap comparison method for ranges has the advantage of being conceptually simple: if the two ranges overlap, then the results are consistent; if they do not, then the results are not consistent. This is a heuristic that students can grasp quickly and that works well for most data.

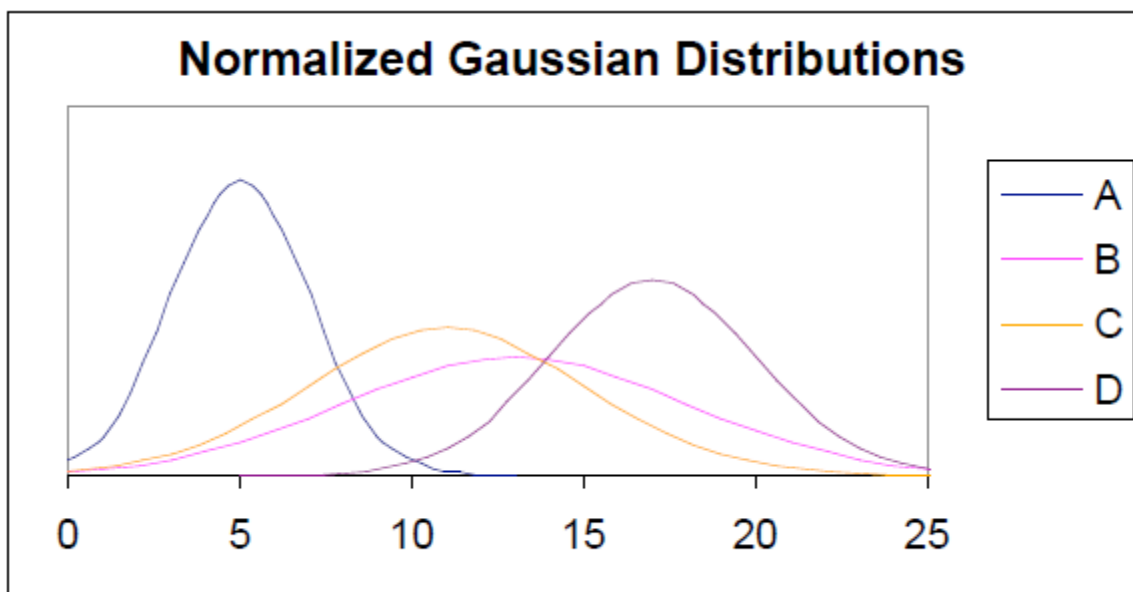
Certainly, there are cases where an overlap approach produces misleading results compared to the more rigorous standard deviation and t-test method. Deardorff [8] examined the pros and cons of the two approaches by examining such an edge case, students to compare the various data ranges shown in Figure 2-3.



**Figure 2-3: Comparison of ranges (reproduced from Deardorff [8]).**

Ranges for four different data sets were shown graphically and numerically, and respondents were instructed to decide which of the data sets were in agreement. Some of the results were obvious: all of the graduate students and 90% of the undergraduates agreed that B & C were consistent, for example; similarly, the consensus was that A and D were not consistent.

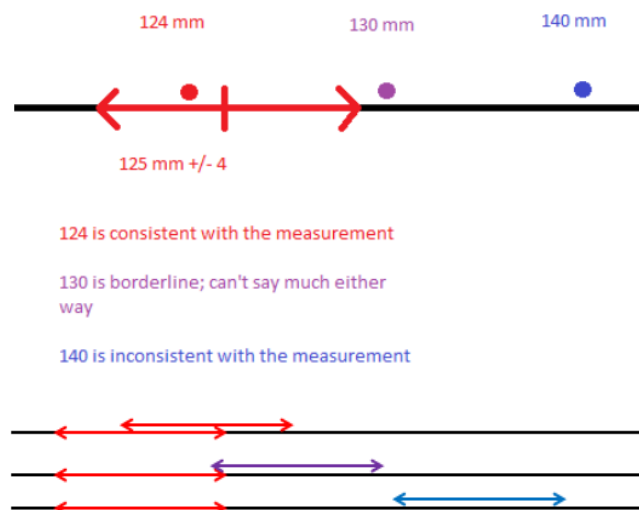
Two of the comparisons are more illustrative of a central difficulty in intuitively evaluating statistics. Around 50% of the students (both graduate and undergraduate) believed that A & C, whose ranges just touch, were in agreement; but 0% of the students in both groups believed that A and B, whose ranges barely do not overlap, were in agreement. Why should shifting a range by a small fraction of a standard deviation produce such a difference in responses? When the ranges are viewed properly as normalized Gaussian distributions, this distinction becomes even more murky, as seen in Figure 2-4.



**Figure 2-4: Gaussians of the ranges shown in Figure 2-3 (reproduced from Deardorff [8]).**

Crucially, this presentation was not shown to the respondents. It shows graphically the statistically meaningless distinction between a range being one standard deviation away as opposed to slightly more than one standard deviation away. Although curves B and C are almost equally likely to be consistent with A when performing a t-test (18% and 23%, respectively), they were intuitively treated very differently by both students and teaching assistants. This is because A and C are consistent according to the informal (but useful) “overlap test”, since their ranges touch; A and B are not, because their ranges do not. Even though the overlap test doesn’t work well in these marginal situations, it does accord with the more precise statistical tests in most cases, as in the non-agreement between A and D. Deardorff advised that if the more advanced statistics were inappropriate for a given class (as in algebra-based introductory physics), overlap is the preferable choice.

The overlap method used in 1-Series was modified slightly to attempt to address these edge cases. Three types of comparisons were possible: consistency, when the ranges strongly overlapped; inconsistency, when the ranges were quite far apart; and borderline, when the ranges just barely overlapped or just barely did not, where nothing meaningful can be said. This is a slight modification of the suggestion for non-statistically oriented classes suggested by Deardorff, as it includes the intermediate “not sure” option, where those statistical methods would normally be required to describe the results properly. No quantitative distinction was made between these types of result; rather, a student’s score on the lab write-up depended on how well she could argue what her data was actually saying. Figure 2-5 was included as a part of an uncertainty and error document to show graphically how to compare different data sets (see Appendix i for the full document).



**Figure 2-5: Comparing a range to a single data point (above) and to another range (below).**

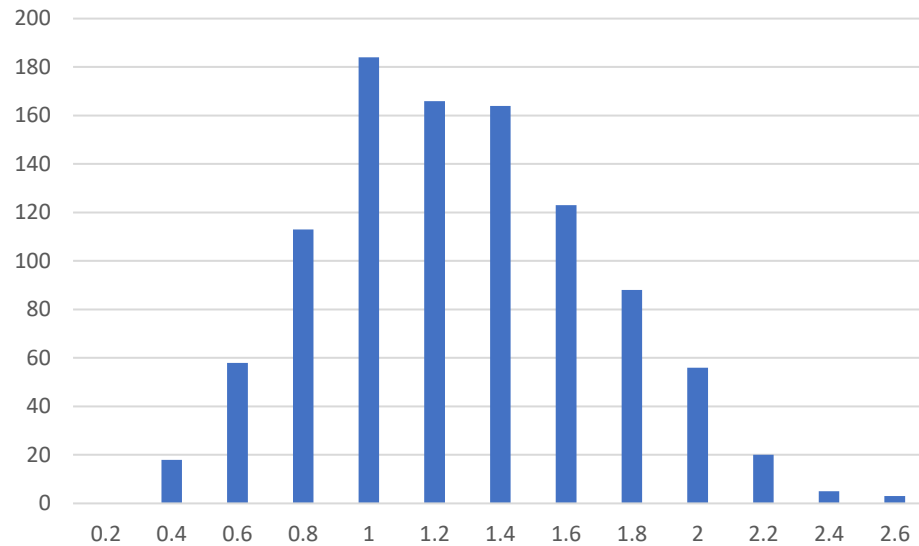
How does this method compare to the more rigorous use of confidence intervals based on multiples of the standard deviation? In other words, how large is the max-min uncertainty defined above for a small number of repeated measurements? To investigate this, random normally distributed data was generated in five measurement sets, simulating lab data. Assuming a normally distributed quantity, the uncertainty for five measurements divided by the standard deviation is shown in Graph 2-2 for 1000 sets of measurements. 89% of the data sets had an uncertainty between 0.8 and 2.0 standard deviations. This places the confidence interval of this uncertainty procedure between the 67% of one sigma and 95% of two sigma; for five measurements, then, there is some uncertainty about the edges of the range<sup>3</sup>. Consequently, it was important to emphasize that the edges of the student-measured range not be taken as absolute, necessitating the middle “no conclusion possible” conclusion shown in Figure 2-5. This is discussed further in description of survey results in Chapter 4.

This approach also does not follow more careful procedure of defining a null hypothesis and either rejecting that hypothesis or not rejecting it. Strictly speaking, there is never any red “agreement” category in science, only a statement that it is not possible to distinguish between the two data sets, or between one data set and the prediction. While crucial for science, this distinction would best be taught alongside the more sophisticated statistical concepts (i.e. standard deviation, confidence levels, and various tests of agreement) in a more advanced lab class. Still, the hope is the simplified versions of these ideas (uncertainty, range, overlap of ranges) will map to the more detailed and rigorous ideas when those students continue their scientific training in subsequent upper division lab classes in their majors, or in actual scientific research.

---

<sup>3</sup> The three-measurement data sets were noticeably more skewed to the low ratio side; 33% of these had uncertainties less than 0.6 standard deviations

**Graph 2-1: Histogram of max-min uncertainty divided by standard deviation for 1000 randomly generated Gaussian data sets of 5 measurements.**



This method does, however, encourage students to accurately report their data, even if it does not agree with their expectations. Students are graded on correctly judging what their data shows, not whether the data itself is “right” or “wrong”. This spread-based agreement approach provides an objective way (with some edge cases) of evaluating any given data set, allowing teaching assistants to grade on how well students interpreted their data, not on how well they collected it. Since developing experimental precision for physics experiments is not a priority for a biology student, these students aren’t assessed on how well they have gathered their data, except in an informal, formative way while they are actually gathering it in class. Practically speaking, students often feel uncomfortable about reporting “incorrect” results (i.e., results that do not support an expected theoretical result), especially at the beginning of the quarter. With proper encouragement and comments from instructors, students begin to understand that their role in writing a conclusion is to report and evaluate their experiment, not to distort or apologize for it.

See Chapter 5 for more discussion on how these goals can change for a field, like chemistry, where lab technique is more important.

This method also elides the difference between accuracy and precision, which is a simple if meaningful distinction between systematic and statistical error that does not rely on mathematical complexity. Defining an uncertainty based on a range of measured values is certainly representing the precision of the measurement, not the accuracy. For most experiments in 1-Series, systematic error is a minor factor compared to the statistical error and is consequently ignored. In a few cases, though, systematic errors in the apparatus are the dominant effect. For example, one lab included a tug-of-war experiment using force sensors connected by a string, meant to show Newton's 3<sup>rd</sup> law in action (with both force sensors recording the same value). In practice, because the force sensors only measure force in a single direction, slight tilts in the angle of the tugging caused the sensors to read different values. The difference between the two force meters was compared to the (much smaller) statistical error, showing that the angle effect was the most significant reason that the two force sensors did not read the same result. This comparison was more about accuracy than precision, since the 3<sup>rd</sup> law was being used as an objective standard of comparison. In general, then, accuracy was presented as a result of the comparison process, whereas precision was presented as the experimental uncertainty itself.

From a laboratory design perspective, each procedure should have a measurable and consequential uncertainty, to allow students to estimate, measure, and propagate it. Ironically, this means designers should keep the experiments flawed, instead of always working to minimize their uncertainties. For example, a lab technician at UCSD offered to fix the aforementioned force sensor string angle uncertainty by replacing the connecting string with a solid aluminum bracket, eliminating the disagreement between the two force sensors. While this would produce a more



precise Newton's 3<sup>rd</sup> Law experiment, the uncertainty analysis of the lab would suffer, since the interesting and noticeable systematic would vanish. Ultra-high precision experiments (where a measurement device is too imprecise to notice the underlying statistical or systematic uncertainties) don't allow students to measure a real uncertainty, and can deny them practice with the supplementary techniques, like comparing data and propagating uncertainties.

#### **4. Treatment of outliers**

In Chapter 1, Allie's survey item relating to the rejection of outliers was introduced [6] in Figures 1-1 and 1-2. In that item, students had taken six readings of the same distance measurement – 443, 442, 436, 588, 437, and 429 – and discussed whether or not to exclude the outlier value of 588. In Allie's treatment, deciding to reject this outlier because it was far outside of the range of the remaining measurements was considered to be a sign of range reasoning expertise. Still, some accepted the outlier also using a range reasoning approach: "the value 588 mm shows how big the spread of the values are and should be used because that is what the group has measured and should form part of their results."

Interestingly, then, the spread-based reasoning cuts across both the "keep the point" and "drop the point" responses. Both of the spread-based responses are basically correct, in my view: the difference between the two of them is access to the machinery of statistically calculating which outliers should be dropped. Lacking that (as in a non-statistical lab class), one would be justified in keeping the point, and expressing its extreme outlier-ness in the spread and in any commentary about the data. (Couldn't it have been a gross procedural error? Yes, although that would usually be investigated just after taking the data, rather than it being recorded by the students by rote. In

this sense, the probe is a bit counterfactual, as students would more likely discuss point 588 as soon as it happened, dumping it before the rest of the points were even available.)

I don't think rejecting outliers as a part of the mean is a part of the skills hierarchy for non-statistical lab classes at all. As seen above, it has at best an oblique relationship to using spread-based reasoning. Without the statistical tools to properly reject an outlier once all data is collected, it relies on student instinct in rejecting a far-off value. Worse, a focus on rejecting data without statistical tests can encourage students to reject data that they really should keep; for example, data that disagrees with a theoretical result that may or may not be valid under lab conditions. The goal of spread-based teaching should be to emphasize the value in keeping different data points, by showing them the value of using the spread to compare data sets to each other and to reference value; rejecting modestly outlying data points just biases the spread. Ironically, then, I'm in favor of using the TA as an authority source (Allie category F) to identify the grossest sorts of procedural mistakes to be rejected. In this approach, students are being guided to get roughly appropriate data by the TA; at which point, they interpret it themselves.

Retaining outliers doesn't hinder the key procedures themselves (i.e., calculating the mean, determining the spread, comparing results with the spreads), it just limits the applicability of the results. For example, keeping an outlier will expand the spread of the data, leading to more likely "agreement" with reference values and other data sets. Since agreement with other data in a procedural sense is value-neutral (students should not be "rooting" for agreement or disagreement in any comparisons), keeping outliers shouldn't hurt skill development. Anecdotally, in some cases where students have kept moderate but noticeable outliers, this has naturally reduced their confidence in their results as stated in lab conclusions. Since one of the broad, high-level goals is for students to see the connection between quality of results and size of the spread, keeping outliers

might ironically prime students to notice that more, via their decreased confidence in their own results if outliers are included.

For these reasons, students were not encouraged to drop outliers, although gross procedural errors leading to possible outliers were often corrected by the teaching assistants before the final recording of the data. This policy also helped to ensure that students did not accidentally exclude data points that appeared to be outliers to the students, but actually represented the proper physical results! See Chapter 5 for further discussion on how unexpected experimental results were perceived by 1-Series students.

## **5. Propagation of uncertainty**

One of the most important lab tasks is for students to be able to propagate uncertainties. This is a key skill in introductory physics because so many of the crucial quantities in physics are not easily measured. For example: momentum, mechanical energy, rotational inertia, heat capacity, power, etc., can rarely be directly measured at all using standard lab equipment; instead, these experimental values are calculated from more easily measured base quantities (like mass, time, and distance) using physics equations. For such concepts, experimenters also need to be able to determine the uncertainty of the unmeasurable quantities, otherwise there is no way to determine how well they agree with expectations.

In a typical momentum conservation experiment, for instance, gliders are placed on a friction-minimizing air track and undergo collisions to show conservation of the gliders' momenta before and after the collision. These momenta are not directly measured; instead, relying on the

equation  $p = mv$ , the carts are weighed (determining their mass) and their motion is measured using photogates (determining their velocities). Even if students measure the uncertainties in mass and velocity properly and calculate the momenta of the gliders, how can they determine if the momentum was the same before and after the collision? Without a number representing the uncertainty of the momenta themselves, a comparison (whether via overlapping ranges or more rigorous statistical tests) of the two data sets is impossible. The actual experimental goal (determining if the momentum of the two carts is unchanged after the collision) cannot be determined without uncertainties of those quantities, which can lead to tepid lab write-ups, as students do not have the ability to answer the questions they are asked. Propagation of uncertainties is the way of using information about the base quantities to figure out the uncertainty of the derived quantity. This allows conclusions to be drawn about experiments whose key quantities are not directly measurable with the lab equipment.

Rigorous propagation of uncertainties requires measurements of: the base quantities; their standard deviations; the equation that relates the base quantities to the derived quantity; and various partial derivatives of that equation. Consider a function  $q(x, \dots, z)$  of independent variables  $x, \dots, z$  which have uncertainties  $\delta x, \dots, \delta z$ . When the variables are random and independent, they contribute to the overall uncertainty  $\delta q$  as shown in Taylor[12]:

$$\delta q = \sqrt{\left(\frac{\partial q}{\partial x} \delta x\right)^2 + \dots + \left(\frac{\partial q}{\partial z} \delta z\right)^2} \quad (1)$$

Although these partial differential equations can be replaced with a set of rules governing different circumstance (e.g., a rule for sums, a rule for products), this still requires a significant amount of overhead to teach these rules which wasn't readily available in the resuscitation-less 1-Series lab instruction.

How are propagations of uncertainty possible using only ranges and without any calculus? Students in 1-Series were responsible for calculating the maximum possible error for the derived quantity by calculating with the extrema of their ranges; for instance, if an experiment required a calculation of the uncertainty in resistance of a light bulb given measurements in voltage and current, the minimum resistance would correspond to the minimum change in voltage and the maximum change in current. From a statistical standpoint, of course, this is the worst case scenario for the correlation of random uncertainties: they all act in concert to increase the total uncertainty. Thus, this procedure represents a very conservative estimate of the derived uncertainty. This is shown in detail in Figure 2-7, which is part of the supplemental document provided to students describing the class's treatment of uncertainty seen in Appendix i.

- Solve for the variable
  - $\Delta V = I * R \rightarrow R = \frac{\Delta V}{I}$
- Using the uncertainties of the measured variables, determine the largest possible value of the new variable
  - $R = \frac{\Delta V \pm \delta V}{I \pm \delta I} \rightarrow R_{max} = \frac{\Delta V + \delta V}{I - \delta I}$
- Repeat the previous step for the smallest possible value
  - $R = \frac{\Delta V \pm \delta V}{I \pm \delta I} \rightarrow R_{min} = \frac{\Delta V - \delta V}{I + \delta I}$
- Take the difference between the maximum and minimum values and divide by 2 to determine the uncertainty
  - $\frac{R_{max} - R_{min}}{2} = \delta R$
- Quote your final answer as variable +/- uncertainty
  - $R \pm \delta R$

**Figure 2-6: Excerpt of uncertainty document showing the how to propagate uncertainties in the case of Ohm's Law.**

This procedure relies on the same quantities as the more rigorous procedure: measured quantities and their uncertainties and the equation relating the measured and derived quantities to

each other. The most challenging part of the procedure is assigning the minus or plus signs to the added uncertainties, since that requires an understanding of how the measured and derived quantities are related to each other. Still, most quantities in introductory physics are directly proportional or inversely proportional to each other; students quickly grasp the pattern of assigning these signs. Moreover, students learning the procedure can “experiment” with the signs, assigning them in all possible combinations to determine the extrema. For example, continuing the example of Ohm’s Law equation  $R = V/I$ , there are four possible combinations of signs:  $(V + \delta V)/(I + \delta I)$  ;  $(V - \delta V)/(I + \delta I)$  ;  $(V + \delta V)/(I - \delta I)$  ; and  $(V - \delta V)/(I - \delta I)$  . Students can think carefully to realize that only the third equation (with its larger numerator and smaller denominator) yields the maximum value for the resistance, or they can simply try each of the combinations, and compare the numbers together to determine the largest value. Early quarter examples of this procedure only use one or two measured variables to allow students to practice and experiment, helping them to figure out the general principles. By the end of the class, students apply the same procedure to multiplicative equations with four or five variables, or even in significantly more difficult cases like the thin-lens equations. Students begin propagating uncertainties in the conclusion of Lab 4 in 1AL; within a few weeks, students are typically comfortable with the procedure. Care must be taken to avoid too many comparisons, because the procedure can often be time consuming.

This procedure works best if uncertainties are relatively small and each variable appears in the equation only once; 1-Series labs do not use more complex equations like the black body distribution equation, where the dependence of a variable is complicated. In these cases, it would be difficult to assign the proper signs, to make sure that the uncertainties aren’t working towards each other.

The uncertainties always “line up”; that is, this procedure overestimate the extrema for the propagated error by evaluating the extreme edges for that variables. This is not present in the formal treatment of the propagation of errors, which assumes the uncertainties in different variables are independent and therefore combines them in a Pythagorean way. While this would be a problem for an experimenter trying to make a measurement more precise, in our view the relative ease of calculation far outweighs the statistical imprecision. The goals of the lab pedagogy run opposite to those of a careful experimenter, who is trying to extract the maximum information out of a flawed experiment; instead, for an uncertainty curriculum, the intent is to highlight the measurement uncertainties for the students, so larger fractional uncertainties are generally advantageous<sup>4</sup>.

While propagated uncertainties often scale multiplicatively with measured uncertainty, lab experiments with sensitive behavior to initial conditions often result in much larger uncertainties than would be naively supposed. This isn’t a negative side effect of the simplified uncertainty procedure, but a real effect based on the underlying equations that can be observed in the lab. The best example of this is in thin-lens optics: if the object is placed at exactly the focal length of the lens, the final image will appear “at infinity”, meaning the final light rays are parallel. This is a special divergent point in the equations; if the object is instead placed within a millimeter or two

---

<sup>4</sup> This “lining up” can substantially increase the uncertainty of the derived quantity when the equation used is a subtraction, especially with small fractional uncertainties. For example, consider  $C = A - B$ , where  $A$  is  $6 \pm 1$  and  $B$  is  $5 \pm 1$ . The largest value of  $C$  is then  $(A+\delta A)-(B-\delta B) = (6+1) - (5-1) = 3$ , and its lowest value is  $(6-1) - (5+1) = -1$ . The uncertainty of  $C$  in this case is  $\pm 2$ , double what the uncertainty for  $A$  or  $B$ . Worse, it represents a very large fractional error in  $C$ . Subsequent additions or subtractions will linearly add to the uncertainty (or multiplicatively add to the fractional uncertainty), leading to a ballooning uncertainty which can be far larger than the best estimate. Fortunately, virtually every equation in introductory physics is a simple multiplicative equation, which ignores these issues. A significant exception is the thin-lens equation, which adds (and subtracts) inverses of length measurements, as discussed in the main text.

of the focal length of the lens, the resulting focused image will appear several meters away from the experimental setup. In the 1-Series version of this lab, students project these images on a solid, moveable piece of plastic, measuring where the image is sharpest to measure the distance from the lens to the image. When the object is near the focal point, this image extends all the way across the lab, even though the lab components are grouped closely together on the tabletop. Students have great difficulty in measuring just where the image is sharpest, since the uncertainty of the image distance is so significant; with the experimental equipment at hand, this “most focused” position can only be measured perhaps to the nearest meter. Beautifully, this manifests itself also when students do the analysis of uncertainties, as the propagated error in the object distance also works out to be  $\pm 1$  meter or so. This is the best case scenario, where a seeming experimental failure turns out to be a consequence of the underlying equations, which the students can confirm during their analysis.

### **Structural changes**

The changes in the method of recording and using measurement uncertainty were dramatic from a content perspective. Unlike the engineering series, 1-Series instruction had no provision for a lecture hour for the laboratory to teach about statistical or other measurement issues. Furthermore, it was not practical to discuss these issues in the lecture for the non-lab portion of the course, since those separate classes were taught by a variety of different instructors. And although a detailed guide to the new system of uncertainty was provided (see Appendix I), it was more useful as a reference than as a practice tool.



Consequently, students needed to learn the uncertainty procedure during the lab classes themselves. This necessitated changes in not only course design and assessment, but also in how the TAs were trained and supported by the TA coordinators. Moreover, these structural changes were also used to help address the most egregious pedagogical issues that already existed in a course consisting of mostly first-time, untrained teachers teaching largely unrevised cookbook labs. While the problems discussed in this section are anecdotal and not measured in any quantitative way, they represent both my experience as a TA coordinator and that of other TA coordinators. These issues and the structural changes used to address them are the subject of the rest of the chapter.

## **1. Learning goals**

Lab manuals, as communal documents, can often drift from their intended focus over time. As text is changed and different activities are inserted, whatever original coherence existed in the document can slowly be lost over time. Any lab reform effort that works with existing lab procedures will grapple with this issue: what is the point of this lab anyway?

The first steps towards this study worked with pre-existing lab manuals as is, with new tasks added around the margins such as lab conclusion prompts, checkpoint questions, etc. At UCSD, lab manuals are printed a year at a time; after a year of experimenting with supplemental documents, it was time to decide whether the lab procedures themselves should be modified to better compliment the uncertainty goals. See Appendix II for an example of a lab written after that year.

Learning goals – short student-centered statements that allow students to determine how well they’ve learned something – were a useful tool in codifying what parts of the lab had a purpose and which did not. If these goals are written with a Bloom’s taxonomy verb, well-constructed learning goals communicate not just what material is to be learned, but how to determine when one has learned it [19]. Although the learning goals were mostly written to help edit lab manuals, they were printed in the new manuals to allow the students to self-assess. To ensure that students read the learning goals for each lab, students were asked to paraphrase one of the goals on some lab quizzes. Learning goals were inserted in the lab manuals near the assigned pre-lab reading, to further associate the goals with preparing for the lab.

- Connect a multimeter to read either voltage or current in some part of a circuit.
- Describe what happens on the plates of a capacitor as it charges, and how it affects other parts of the circuit.
- Use capacitor adding rules to determine how long it takes for light bulbs in various circuits to go out.

**Figure 2-7: Learning goals from the first circuits lab in 1BL.**

The learning goals generally did not touch on uncertainty, but were focused on the key tasks, concepts and how to use the equations of that lab. When revising the lab manuals, any lab that had too many (or too diverse) learning goals was flagged for additional editing. A lab manual with too many distinct learning goals was probably overly complex conceptually or at least trying to do too many things at once.

## 2. Pre-lab Questions

Pre-lab questions began as physics homework problems that students needed to complete before attending a lab. Because the lab and lecture classes can be out of synch for the students, pre-lab questions should allow students who have yet to learn about the topic of the lab to catch up with those that are experienced in that physics already. The problems, however, were often too difficult for students who had little experience in the topic; worse, some of the lab questions were originally designed as post-lab summative assessment but were later assigned as pre-lab tasks! Moreover, lab reading assignments were often vague (e.g. “read chapter 13”) and at times only glancingly appropriate to the lab in question. And because the lab sections and lecture sections were mixed, the same lab class could include students who had covered the material in the lecture and those that had not, making the teaching even more challenging.

This led to the following negative outcomes:

- Students often could not successfully complete the pre-lab questions, especially if the material hadn’t been covered, because of its difficulty.
- Since pre-lab questions were worth 5 points, 25% of the lab score was determined by occasionally very difficult problems that unprepared students shouldn’t have been expected to complete.
- This higher pressure assessment led to more students copying one another’s pre-lab work, contrary to the academic integrity policy of the course.
- Pre-reading was rarely completed, leading to students not being familiar with key terms that they needed for the lab.

- Students whose lecture sections had yet to reach the topic were disadvantaged compared to those that had spent time on the material in their own lectures.
- On the “difficult question” weeks, office hours were dominated by students needing help on pre-lab questions, leaving less time for the in-principle more difficult post-lab assignments.

What purpose should pre-lab questions be serving? Taking inspiration from Crouch and Mazur [20], these questions are best viewed as formative assessment, encouraging students to do the pre-reading to prepare them for the lab. Consequently, the questions and pre-reading were changed in the following ways:

- Pre-reading was limited to key terms, small relevant sections of text, and diagrams directly relevant to the lab. In cases where the textbook didn’t cover the necessary grounding, computer simulations (such as pHETs [21]) were used to encourage play.
- Difficult or lengthy pre-lab questions were removed entirely.
- Point values for the pre-lab questions were reduced to 3 points instead of 5 points (out of twenty)

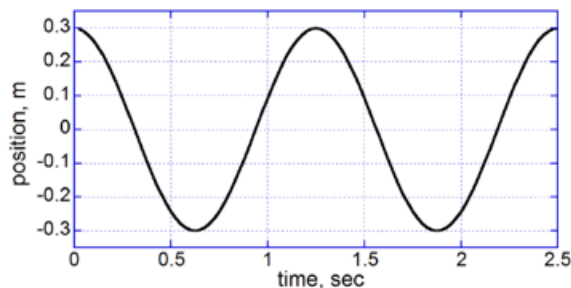
In an effort to make pre-reading less overwhelming, it was presented as a list of bullet points that targeted very specific portions of the text. Often, these reading sections would be relevant to one or more of the learning goals of the lab, further supporting the utility of those goals (see below). In cases where the amount of new material was too large (for example: the transition from Snell’s law to single lens refraction), the University of Colorado’s online physics simulations (pHETs) were used to allow students to play with the ideas of the lab beforehand. Finally, the

fewer points assigned and overall easier (and more fair) content served to help decrease academic integrity violations.

### 3. Lab Quizzes

Although pre-lab questions are a good way of encouraging students to prepare for the lab, they are only partially an individual activity. Students can work together on the questions as long as they make sure to separate before writing up their responses. The lab quiz served as a separate check to ensure that individual students had prepared for the lab.

1. In the equation  $A \cdot \sin(Bt + C) + D$ , how can you estimate the parameters  $A$  and  $B$  in a given test without performing measurements of actual oscillations using the position sensor? (2 points)



2. For oscillations of an object on a spring shown in the plot of, what is the frequency? (1 point)

Draw a plot of acceleration as a function of time.

Label the maximum value of acceleration. (2 points).

1. Define the phrase “simple harmonic motion.” Why does a mass on a spring exhibit simple harmonic motion? Explain. (3 points)

2. What is the angular frequency of a mass on a spring determined by? (2 points)

**Figure 2-8: Example pre-lab quiz questions before changes (above) and after changes (below).**

The lab quiz occurred at the beginning of the lab, just after class had started. The quiz ideally took around ten minutes of the lab, covering both lab procedural details and simplified versions of the pre-lab questions. Quiz questions are one of the few direct responsibilities of the course instructor, and so can change from quarter to quarter. As an assessment that relies on the students internalizing the ideas of the pre-lab questions as well as some procedural details of the lab, a test environment that prohibits notes is essential.

Before the modifications, lab quizzes had the following issues:

- Quiz questions based on the difficult pre-lab questions were too challenging, especially under a time constraint. Sometimes, they covered novel physics from the lab.
- Teaching assistants responded to quiz difficulty by extending the amount of time for the quiz, cutting into the time for the lab proper.
- Because there was no specific quiz protocol, teaching assistants overall were not proctoring the quiz. This led to frequent cheating, especially since students were allowed to form their own lab groups that stayed the same each week
- After some students completed the quiz, the quiz environment became more chaotic, as they began to take out their lab manuals and turn in their lab work. Since some quiz questions could be answered simply by reading the lab, students who took their quizzes slowly were sometimes able to examine others' lab manuals.

In Bertram et al.'s study [22] on academic integrity violations in the 1-Series, some interviewed students stated that they had seen or heard about TAs turning a blind eye to cheating (consulting lab manuals and talking) in the lab quizzes. This sort of activity can lead to more

cheating, as students are more likely to cheat if they think others are doing so. While the problems with the test-taking environment were not endemic, TA training did not detail how to ensure the environment was fair, leading to a wide variety of outcomes. It was clear that a universal quiz procedure was needed that ensured that each student was fairly assessed. The initial idea for this procedure came from a teaching assistant (Chris Murphy) who had his students walk outside after completing their quiz. In this way, everyone in the lab room at any given time would be taking the quiz, limiting the possibility for looking at lab manuals during the test.

The eventual quiz protocol we developed and trained TAs with was as follows:

- Student seating is randomized as they enter class (usually using playing cards), assigning them to work with an unpredictable group each week.
- TAs begin the quiz on the hour. Students that have yet to turn in their work are instructed to put it away for now.
- Desks are cleared of lab manuals before quizzes are handed out.
- TAs walk around during the quiz, making sure students are working alone.
- Students who complete the quiz turn the quiz in, then walk outside for the remaining duration of the quiz.
- Once all students are outside, the teaching assistant announces the end of the quiz, allowing students to reenter class and proceed with the lab.

Moreover, TA coordinators visited each lab class within ten minutes of it starting, to make sure the quiz process is being followed. (TA coordinators were often already observing classes at this time, since they were responsible for teaching the lab section if the teaching assistant was late or absent)

These changes made the test environment significantly more fair and stable.

Quiz content itself was less of a concern. With less difficult pre-lab questions, the quiz questions based on them also became less challenging, and more appropriate for the short test period. Procedural questions focused on broad ideas in the lab, rather than detailed and easily confused equipment questions. This helped to ensure that students who had completed the pre-reading received the appropriate credit for their preparation.

#### **4. Checkpoint questions**

Aside from weekly lab training, which gave some general advice on running the lab, the teaching of the lab itself was left up to the TAs. While TA coordinators were occasionally present to help, they were typically responsible for observing between 2-4 labs at once, limiting the amount of time for any one lab. Consequently, teaching assistants were responsible for running the quiz, managing students' time, answering student questions, and addressing procedural difficulties. The best teaching assistants managed to juggle all these responsibilities while also asking their own questions of the students, helping them to understand the material. The typical teaching assistant for the lab class was a first-year student teaching their very first class, so the lack of instructional support and guidance led to the following negative outcomes:

- Students were not able to complete all labs.
- Some teaching assistants took a very passive approach to teaching, sitting down and often not initiating conversations with students



- Because of this, equipment problems were sometimes not discovered until they had taken up too much of the students' time.
- Proactive or friendly student groups could monopolize the TA's time, leaving less attention for groups that were more passive or independent
- Students sometimes did not record crucial data they needed for their lab report.

The challenge was to find a way to address these issues without burdening teaching assistants with too many management duties. The priority was that students would get regular progress checks throughout the lab, regardless of whether they or the teaching assistant were passive about talking to each other.

“Checkpoint questions” were developed: three in-lab questions per lab that were designed as points of forced interaction between the teaching assistant and the lab group. To get full in-lab points, a student would need to get “checked off” by their TA three times in the class, once for each checkpoint question. Questions were written on the shared class whiteboard, and consisted sometimes of procedural check-offs and other times of conceptual discussions. Because each step of the lab manual had an alphanumeric coding, it was clear at a glance when to work on each checkpoint question.

Crucially, these questions empowered student groups to force their TA to talk to them. Since the students certainly wanted the participation points for completing the checkpoint questions, they proactively engaged their teaching assistants, even those that were shy, passive, or distracted. Checkpoint questions are ideal for safety issues, since they require students to check in before moving on to power on a potentially overheating circuit, for example. Beyond that, the

checkpoint questions established a list of acceptable questions that students could ask about even if they were generally unwilling to expose their lack of knowledge.

A6: How does the bulb need to be connected to the battery in order to light up? Answer this without using the phrase “complete circuit.”

F8: Demonstrate how you measured voltages (with uncertainties) at different points around the circuit in parts F5-6. Demonstrate how you can measure the voltage “used up” by the light bulb.

F14: Show your TA the diagram you drew for the two-battery circuit, and discuss the differences in the measured potential differences compared to the single-battery circuit. Why must the batteries be connected “+” side to “-” side? What would be different if they were hooked up “+” to “+” or “-” to “-”?

**Figure 2-9: Checkpoint questions from 1BL Lab 2 (introduction to circuits).**

Before the checkpoint question system, it was often very difficult to engage students in material, especially uncertainty, that was not explicitly discussed in the lab manual. Verbal reminders are only so effective and can easily be misheard once and written down incorrectly. Moreover, less experienced teaching assistants had been less likely to be able to gather the attention of the entire classroom, especially consisting of students concentrating on their labs; this also limited the effectiveness of verbal full-class announcements. Students would often forget to record the uncertainty for a measurement, which led to their lab-write-ups being essentially impossible to complete. Not only would a checkpoint question about recording uncertainty remind students to actually take the measurement before moving on, the teaching assistant could individually check each student’s manual to ensure that it was correctly recorded and labeled.

Another possible system of making sure that students have recorded their data is checking off lab notebooks at the end of class. Checkpoint questions were preferable in this case, because

the checking is done during the lab, while the students are still investigating the subject. If the student group has misinterpreted the instructions, for example, and not taken the uncertainty in the correct way, the teaching assistant can give them further instruction and the time to record the data correctly. End of lab checks, by definition, do not allow for redress of student mistakes, since there is no further time in the lab class to correct them. This is also why the lab conclusion prompt is announced with thirty minutes to go in the class, rather than right at the end (See lab conclusions, below).

Despite their overall utility, checkpoint questions do present significant challenges for time management. Each question must be checked off separately for each lab group; with the typical six lab groups in 1-Series classrooms, that is eighteen discussion points in each two-hour lab. Great care must be taken to make sure some of the checkpoint questions are quick procedural checks, to limit the burden on both TAs and students. If checkpoint questions are too frequent and complex, lab groups will spend significant time waiting for the attention of a TA, instead of focusing on the lab. These bottlenecks can potentially derail an otherwise functional lab. The current number of checkpoint questions were designed for a two-TA lab, and UCSD has the additional benefit of the roving TA coordinators to fill gaps. For a single teaching assistant and no rovers, decreasing the number of checkpoint questions to two would be appropriate.

## **5. Lab conclusion structure**

After each lab class is completed, students were required to write a “lab conclusion”, a 1-page write-up to be turned in during the next lab class. Before the lab changes, the content and detail of these lab conclusions were the responsibility of the teaching assistant to assign, and so

varied greatly. While that might be a fine choice if the lab instructors were faculty or experienced teaching assistants, it was a bad match for many of the 1-Series teaching assistants, many of whom were teaching their very first class. These were some of the more common failure states:

- Teaching assistant training for the lab was minimal, so assigned conclusion questions may have little connection to the most important ideas of the lab.
- Some teaching assistants gave such instructions as “write your conclusion on part A of the lab” with no further detail provided, leaving confused students and TA coordinators doing triage in office hours.
- Some TAs were more specific, but didn’t record what they asked, leading to frequent misinterpretations by their students, or the need for follow-up questions via email.
- In the worst cases, teaching assistants might ask students to prove something that was impossible, having misunderstood the physics behind the lab.
- Because of all this variability, students attending different sections were asked to do very different tasks for the same grade.
- Uncertainty was rarely a significant feature of lab conclusions.

Rather than continuing to allow the teaching assistants to determine the structure of the lab conclusion, a single prompt for each lab was written in advance by the TA coordinators. These standard conclusion prompts were central to the effort to modify the lab classes to include uncertainty. With no lab final exam, lab conclusions were the lone piece of truly summative assessment in the class; consequently, it was crucial that uncertainty be a feature of each and every lab conclusion. Before the modification to the lab class, measurement uncertainty was referenced only in speculation about a nebulous “error”, if at all. Standardizing the conclusion prompts to

cover each lab section had the side benefit of making the lab sections much more universal, in addition to ensuring that uncertainty would be assessed in the class.

Conclusion prompts were not made available to the students before the lab class; instead, they were written on the class whiteboard but concealed until around thirty minutes were left in the two-hour laboratory. This meant that students needed to work on each part of the lab with equal attention, since the write-up could be on any portion of the lab. With a half hour to go, the teaching assistant revealed the prompt, allowing the students to decide whether they were prepared to write a subsequent lab conclusion on that topic. This choice was made to allow students time to perform basic triage on their lab, especially to ensure that they had time to gather the most crucial data and uncertainty for the prompt. Significant time-management problems arose if the teaching assistant revealed the conclusion too late, as students would not have enough time to self-correct. One of the crucial priorities of the TA coordinators was to check each classroom with thirty or forty minutes left, to make sure the conclusion was available in time.

The conclusion prompts were formatted in three sections: conceptual, data, and uncertainty. Since the lab conclusion was the centerpiece of the lab, the point values for it was increased from 5 points to 7 points (out of 20). Usually, three of these points were assigned to the uncertainty section. The overall format follows:

1. Conceptual

General physics questions that related to ideas used later in the conclusion. These questions were intended to be clear to students that had finished the lab.

## 2. Data

Only a reporting of the asked-for data, not the results or conclusions drawn from them. Since any comparison of data (to other data or a reference value) requires the use of some sort of measurement uncertainty, such comparisons are only asked for in the uncertainty section.

## 3. Uncertainty

Always begins with a numerical value (including units) for the uncertainty of at least one directly measured quantity in the lab. Often followed by data comparisons or propagated uncertainties that rely on the measured uncertainty. The emphasis was concrete comparisons and calculations, not commentary on whether the experiment was successful or not.

Lab conclusions consisted of answering the relevant conceptual questions, reporting the asked-for data, and answering specific questions about the data using the uncertainty procedures described earlier this chapter. Some examples of these prompts can be found in Figures 2-11 and 2-12. In particular, students were not responsible for “lab report writing”, a typical lab write-up format where students write a simulated version of a scientific paper describing their experiment and its results. While learning to write a scientific paper is an important part of any scientific education (and lab is a great place to start), it is important that it is taught according to the norms of each individual scientific field. Consequently, 1-Series students as future biologists should be taught how to write biology papers in a future biology lab class. See Chapter 5 for more discussion on what tasks should be taught where for non-major physics students. The choice to focus on specific questions also allowed for shorter lab write-ups (1 page was the general guideline) to allow for a less taxing schedule for both TAs and students.

## 6. Lab conclusion uncertainty content

This choice of format allowed the conclusions to be primarily centered on uncertainty, rather than also having to address general scientific paper concerns like how to write an abstract, etc. Because of the dearth of additional lab time, the lab conclusions were the main area where students were able to practice and learn the uncertainty procedures of the class. For convenience, the main goals for students in the lab are repeated below:

- Calculate or estimate the measurement uncertainty for *each* measured quantity
- Interpret this uncertainty as a measure of variability of the quantity
- Judge the consistency of multiple data sets (or a single data set with a theoretical quantity) using this variability
- Propagate uncertainties to facilitate meaningful comparisons, if necessary

Aside from the actual value of uncertainty (which was sometimes assessed by the teaching assistant in class as a checkpoint question), each of the other skills was only assessed by the TA reading a post-lab conclusion writeup for each student. With only ten labs in each semester (and thus only nine lab conclusions), each needed to carry its weight in terms of helping to teach and assess uncertainty.

During the planning stages of this study, an initial task was determining how these uncertainties practices should be deployed over the course of a quarter. A major inspiration was Kung's *Scientific Community Labs*, a full course designed from the ground up to teach students about uncertainty and to help them communicate as scientists.

Her goal in developing this course was to create “a natural scientific community for developing students’ ideas about measurement and uncertainty [23].” Rather than being a traditional lab course, students were responsible for designing an experiment, performing it, and

providing an analysis of an open-ended question. These provide an example of what Arons called “guided inquiry” labs, as a natural medium between procedurally focused cookbook labs and pure inquiry labs where goals may remain unspecified [2].

Goals	1	2	3	4	5	6	7	8	9	10
Predictive vs. Descriptive										
Measuring Time	Reaction time to catch a ruler.									
Multiple Measurements		Does mass or length change period of pendulum?	How does period depend on length?							
Range Overlap				What affects accel. of a rolling						
Stacking										
Systematic or Random Mech.										
Internal vs. external										
Representations										
Peak overlap, low prob. data					Are the cans the same or different?					
Minimize external variation						Does friction depend on contact area?				
Range Propagation							What size target?			
Predict certainty								Release height for ball to go around loop. Measure g.		
Expand theory									Period of mass oscillating on a massful	

**Figure 2-10: Schedule of uncertainty concepts from Kung’s Scientific Community Labs [5].**



Kung's course was designed not only to teach the techniques of calculating and measuring uncertainty, but also to provide a context where the use of these skills can be demonstrated to the students. The inclusion of class presentations and discussions (where students critique each other's experimental design choices) helped to solidify the lab group as existing in a larger scientific community. This helps to pull students out of the "what arbitrary things is my TA asking me to do" traditional lab frame, to allow a more nuanced conversation on how best to design and interpret scientific experiments. To further support this focus on community building, the lecture and discussion parts of the class included small group work on worksheets and novel homework problems more focused on communicating (e.g., essay questions, representation-translation questions, etc.) than the more traditional calculation practice [5].

This course directly addresses each of the broad lab goals spelled out in Chapter 1. Ideally, this project would have entailed implementing this system in a more wide-scale setting. Unfortunately, there are significant difficulties in implementing such an elaborate structure into the more traditional class context of UCSD. *Scientific Community Labs* really is an integrated lecture / discussion / laboratory course structure, but at UCSD students from a particular lecture can be assigned to any lab class; in other words, there was not a way to change a single lecture section and a small portion of the lab to match.

In many places, including UCSD, the progression of lab experiments matches an implied lecture topic order, rather than in *SCL* where it is designed to explore uncertainty concepts. Traditional labs are meant to complement the lecture class, and so they need to keep pace with the lecture on a topic by topic basis; it is a problem if the lab is discussing a topic far before or far after the lecture class. Strikingly, the final lab topics in *SCL* concern themselves with measuring the local gravitational acceleration "g" in weeks 8-10, months after a typical lecture class would have

discussed that topic, which often comes near the beginning of the quarter in kinematics. Kung's class really is a lab course designed to teach uncertainty and thus the practice of science, rather than the topic of (previously-studied) physics per se. Consequently, Kung's labs included substantial guided group discussions with TAs, and in-lab group presentations; because the labs were designed from scratch, these valuable tasks could be baked in to the existing lab time.

The difficulty in implementing such significant changes transcends not just the UCSD physics department's particular course structure. In physics departments where faculty members teach lab sections, for example, significant changes in lab instruction may affect several instructors. There is a certain inertia that comes with cookbook labs, since an instructor can get the basic idea of how to run the lab by reading the procedure and trying out the equipment. Changes are slow, and broad philosophical changes require significant buy-in from instructors, faculty chair, and also possibly other departments for service courses like 1-Series.

While it may have been possible to implement something like *SCL* at UCSD for a small number of lab classes as an experiment, the broader goal was to improve the 1-Series instruction of uncertainty in a more sustainable way. In the initial implementation, the lab manuals themselves were kept mostly static; instead, the focus was on developing the supplemental materials (checkpoint questions, lab conclusion prompts, etc.) to help teach uncertainty, and then to train teaching assistants in how to use them. Some of these teaching assistants would become the next generation of TA coordinators, teaching the next class of incoming new graduate students in these methods. See the next section for more information on TA training.

Matching the uncertainty task teaching with the standard lab order (the usual topic progression in the lecture class) was sometimes a challenge. Sometimes, relatively early labs used

complicated equations whose properties were a poor match for the simplified uncertainty system. For example, students for Lab 3 in 1AL balanced a force table by measuring the weights and angles on each side, and figuring out where to place the (unknown) mass that would center the table. Propagating the errors for this situation would have been a mess: the force values are multiplied by the trigonometric functions of the angles, making it conceptually quite difficult to calculate errors in the unknown mass and angle especially for students just learning the procedure.

The early labs focused on the most important part: getting students used to the idea of thinking about and measuring an uncertainty, rather than assuming a first measurement was enough. Luckily, the first lab class in 1AL was already well-designed for a discussion of measurement uncertainty. The procedure instructs students to use rulers to measure various objects in the lab room. Instead of just calculating the standard deviation of the measurements, students were instructed to make the measurements once and estimate the uncertainty of the subsequent measurement. For this first measurement, it was crucial to discuss a measurement with students just when they finished making it, thus talking them through the idea of how to estimate uncertainty.

The lab procedure for Lab 1 instructed students to make a measurement of the entire back wall of the lab; such a measurement is difficult to make because of obstructions at ground level. Most students attempt to measure the size of the wall with retractable tape measurers, which noticeably sag; others use meter sticks which need to be lined up several times, often at different heights. Inevitably, when the question is posed “how precise do you think your measurement was?”, the response from the students is “not very precise”. This provides a perfect opportunity for the teaching assistant to press students to provide numerical estimates of the uncertainty of their measurement, presented as how different they think the real length of the wall could be from

their measurements. Crucially, students usually see very quickly that the uncertainty of measuring the back wall is much larger than the tiny tick marks of the ruler. Students are often uncomfortable with estimating a number for the uncertainty, seeing it as arbitrary or “just making something up”. A key lesson here is that there are many possible estimates that are reasonable, and many that are not.

**Conceptual (2):**

Answer the following questions based on your observations in the lab only. Explain and justify your answers to each. How many types of charge are there? What is the difference between conductors and insulators?

**Data (3):**

Draw the *expected* distribution of top and bottom charges for both pieces of aluminum foil for the following situations for part C2 (foils initially touching):

- The foils are touching each other and the charged PVC pipe is nearby
- The foils have been separated and then the PVC pipe has been moved away

Do the same for the setup in C4 (where the foils were not initially touching) before and after the charged PVC pipe has been moved away.

For C2 and C4, describe the *expected* final charge distributions in words and explain why these two scenarios should result in these distributions.

**Uncertainty(2):**

Report your six measurements for how far apart the tapes needed to be before they interacted strongly. Calculate the average value of that distance and its uncertainty.

**Figure 2-11: Lab 1 conclusion prompt from 1BL.**

The measurement lab ties measurement uncertainty, a new concept, to an activity that students already know how to do (measure objects with a ruler). There are no propagations of

uncertainty to make and, indeed, no direct comparisons of range or agreement. An early-quarter conclusion prompt can be seen in Figure 2-11.

Near the end of a quarter, the conclusion prompts are much more advanced, since they include those basic elements as a matter of course and are focused on more advanced skills. For example, in Lab 6 of 1BL, students needed to make measurements of current and voltage, record their uncertainties, propagate them into resistance uncertainties, and use them to check an assumption (in this case, the lab manual's estimate of the current resistance of a bright bulb).

**Conceptual (2):**

Why does the current flowing into a junction have to equal the current flowing out of a junction? What would happen if this were not true?

Does the junction rule hold at every point in the circuit (e.g. points C or D in the diagram from Part A)? Why?

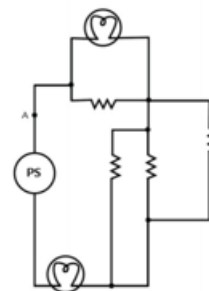
**Data (3):**

Draw the circuit from parts C1-6. Include an ammeter connected at point A and a voltmeter connected across the topmost resistor. Record your measurements (with uncertainty) of the voltage and current through the topmost resistor and the current through point A.

Using only these three measurements, calculate the resistance of the topmost resistor and the resistance of the top light bulb.

**Uncertainty (2):**

Calculate the uncertainty in the resistance of the top light bulb and resistor for the circuit in C1-C6. For the light bulb, are the assumptions from part C4 regarding its resistance correct? Is the resistor's resistance consistent with its labeled value of  $10\ \Omega \pm 5\%$ ?



**Figure 2-12: Late-course conclusion prompt for 1BL, including the circuit diagram provided in the manual.**

Crucially, no weight was placed on the “accuracy” of the final result. In some cases, students found their calculated resistance and uncertainty consistent with the estimate in the lab manual; in other cases, they found them inconsistent. As long as students showed the correct process (measured a plausible uncertainty for the voltmeters and ammeters, calculated the resistance of the bulb correctly using Kirchoff’s rules, propagated the uncertainty of the resistance, and compared it appropriately to the result), the **actual yes-no question of consistency doesn’t matter**. After all, a properly planned, performed, and analyzed experiment might agree with expectations or might not: either result is (or should be) publishable.

## 7. TA training

Because the 1-Series physics lab can serve up to 900 students in a quarter of up to 26 students per lab section, many graduate student teaching assistants are hired to split the work. Most incoming graduate students do not already have a research advisor in their first year, so many of the TAs for the lab classes are first year graduate students, with all that implies: they are taking graduate classes for the first time, with a healthy knowledge of physics but usually no experience teaching outside of one-on-one tutoring. And, of course, they have no experience with the particular laboratory classes at UCSD. Traditionally, TA training consisted of weekly meetings between all of the TAs of a given lab course and the assigned TA coordinator(s). The nominal goal of these meetings was to acquaint the new teaching assistants with the procedure of that week’s lab, giving them some hands-on time with the lab equipment and advice from an experienced lab instructor.

Unfortunately, training meetings were often far too short and surface level, leading to the following issues:

- With typical training sessions lasting under thirty minutes for a two-hour laboratory, there was often not enough time to even discuss each part of the lab setup in detail.
- The TA coordinator running the training often handled the equipment themselves, denying the teaching assistants the chance to play or experiment.
- Lack of experimentation meant that common problems with the equipment could not come up organically, putting pressure on the coordinator to remember each common problem, lest the TAs have no experience with it before teaching.
- Rarely, an elaborate or subtle experimental setup would become disarrayed via normal wear-and-tear and go unfixed, because the TA wasn't aware of this problem; this led to subsequent students doing experiments that could not work as intended.
- With such a limited time, barely any time was spent on non-procedural issues, like common student misconceptions, how to best present the material, and discussion of the conclusion questions.
- Because of the assumption that graduate students would know all of the physics, occasionally TAs would end up teaching a lab with an incorrect idea of how the physics underlying the lab worked, leading to significant downstream issues for their students.

Collected together, these issues meant the typical TA was undertrained and underprepared to teach a new laboratory class. The lax training meant that TAs were much more likely to be unable to handle minor equipment problems, leading often to valuable student-teacher interaction time being

lost to tinkering with the equipment. Training was seen as a time-waster even to some TA coordinators, which further incentivized quick meetings. While many TAs were happy for brief training meetings, some more conscientious teachers found that the training did not provide enough time for them to learn the lab to their own satisfaction; most eventual TA coordinators had put in additional time on their own to learn the labs when they were teaching assistants. In a way, the changes to training meetings were to reproduce the best practices that strong, interested teachers felt were necessary for their own development.

Instead of being an introduction to the lab equipment, the new TA trainings were envisioned as practice in being a student trying to complete the lab. Like students, TAs worked together in small groups, reading the lab procedure in detail. They set up the experiments themselves, taking some data themselves, and working through the checkpoint questions. The TA coordinators served as the “TAs” of this simulated lab session, but with a focus on highlighting common student misinterpretations of both the physics and the procedure. In addition, coordinators provided specific advice on how to teach some of the material (e.g., not using the idea of “resistance” in the first circuits lab, since that concept had not been introduced in the lab curriculum at that point). Coordinators also occasionally prompted TAs to answer student-like questions on the fly, allowing them some practice in considering these questions.

By reading the lab manuals themselves, the TAs were able to find typos and possible misreadings on their own. The small group structure also allowed more experienced TAs to provide their insight to the first year graduate students that were the lions’ share of 1-Series Lab teaching assistants. By doing some amount of data collection on their own, they were able to measure not only the trends the students would be looking for in the lab, but also the amount of measurement uncertainty appropriate for the experiment, when done properly. Without this hands-



on experience (including taking actual data), TAs would lack the heuristics that allow experienced TAs to tell from data whether students have made some sort of procedural error in their collection. By the time the coordinator walks the TAs through the conclusion questions, they have enough of the context of the lab to grasp what they're really about and why they're being asked. While these new TA training sessions were not as long as the 2-hr lab, they were usually in excess of one hour.

The TA training was some of the most challenging work for the TA coordinators. The vast majority of lab TAs were first-year graduate students, who often had little to no classroom teaching experience – and even those who did almost never had a PER background<sup>5</sup>. First-year graduate students need general teaching advice, which was difficult to provide in the context of training them for these labs. Some lessons proved easy for new TAs to apply (make sure you've gained the attention of the whole class before making announcements), and some were more difficult (don't interrupt students while they are explaining something). A portion of teaching assistants felt the training was superfluous, since they “already knew the physics”. Like the lab students themselves, the teaching assistants needed to be kept motivated. Unlike the students, there was no carrot of a final grade to keep them going. TA training was mandatory, and it helped that the department staff took attendance as seriously as the coordinators themselves did.

Despite the difficulties, developing a robust TA training procedure was the only thing that allowed the lab changes to be applied to the entire cohort of 1-Series lab students. This procedure was in principle self-perpetuating to some extent, since some of the students in the training would eventually become TA coordinators in future years and run future sessions of that training. Still,

---

<sup>5</sup>At UCSD, while there was a teaching methods class for graduate students at the time of this study, it was not required and was often taken lightly by the students and their advisors. Moreover, its lessons could still come too late for a graduate student's first teaching assignment, which often happens during their first quarter.

this relies on these future TA coordinators being trained not only to be good teaching assistants, but also good coordinators. New TA coordinators were trained via an informal apprenticeship period (i.e., being matched with a more experienced LTAC for their initial course), including some supplementary materials, but a more formal process similar to TA training was not developed.

## Chapter 3 : Physics attitudes

### Introduction

“Attitudes towards physics” is a PER term describing Likert-scale surveys that are easily deployed to large numbers of students at a time without the need for interviews. Students’ agreement or disagreement with statements -- such as “*Problem solving in physics basically means matching problems with facts or equations and then substituting values to get a number*” or “*When I solve a physics problem, I explicitly think about which physics ideas apply to the problem*” – serve as a tracer for expert-like thinking about physics. While they certainly don’t provide the same level of insight into an individual’s thought process as interviews or problem-solving, attitudes surveys are very quick, require relatively little data analysis, and differentiate between expert and non-expert approaches. In general, people on the expert track in physics (faculty, graduate students, even undergraduate majors) respond to the majority of these survey items as an expert “should”, regardless of level.

The preferred attitudes survey in PER is the CLASS: Colorado Learning Attitudes about Science Study, which was designed to cover overall attitudes towards physics, rather than lab experiences in particular [24]. (After the data was collected for this project, a lab-specific attitudes survey was released that would have likely been a better overall fit for this project [25].) In addition to providing an overall “physics attitudes expertise” score, many of the survey questions are grouped into empirical categories containing correlated sets of questions concerning distinct

aspects of expert attitudes. These categories are: Real World Connection, Personal Interest, Sense Making, Conceptual Connections, Applied Conceptual Understanding, Problem Solving Confidence and Problem Solving Sophistication. Some of the individual survey items fit into multiple classifications, and some don't fit into any, although they still contribute to the overall CLASS score. Because of the careful empirical approach in creating the survey categories, some initially distinct survey items became grouped together; for example, personal interest and effort-based survey items are grouped together under a single category.

Most CLASS survey items (36/42) have an expert consensus, determined by polling physics professors in the original study. From a content knowledge perspective, incoming undergraduate students are not experts. But might they still be experts from an attitudes perspective? Even without the expertise required to solve problems and recognize their key features [26], they might still “enjoy solving physics problems”, believe that “reasoning skills used to understand physics can be helpful to me in my everyday life”, etc. This provides a more affective way of considering physics: do students like physics? Do they see its value? Do they understand its multifaceted approach to solving problems, even if the techniques may be for the moment beyond them?

Thus, surveys like the CLASS provide a complementary way to evaluate student development outside of the much-studied advancement of student content knowledge over the course of a single class. Since experts and non-experts are distinct groups with respect to their attitudes towards physics, one goal of the CLASS was to measure how much a single course could change a student into a more expert-like thinker. CLASS's famous finding was a negative one: typical lecture classes not only find little attitudes progress of the course of a quarter, but also often find the students becoming *less* expert-like. The original CLASS validation study reported that

unless special care was given to student attitudes and beliefs, overall scores tended to deteriorate in post-instruction surveys [27]. Since traditional lectures almost by definition are not explicitly focused on student attitudes towards physics, one would expect students who take several lecture courses in a row would become less expert-like over time.

This leaves an important question: if traditional lecture classes taken individually do not cause student attitudes to become more expert-like, where do these experts come from? Are physics attitudes experts made in undergraduate education at all, or do they already have those attitudes beforehand? This second idea would suggest that the undergraduate to graduate school journey is more about weeding out those students that don't have expert attitudes towards physics, rather than actively changing the attitudes of individual student. Caution is required in the interpretation of these results: the goal of attitudes analysis cannot be provide a numeric justification for gatekeeping. See Chapter 5 for more discussion of approaches to attitudes surveys.

Gire used CLASS to examine physics majors, in an attempt to determine if attitudes-based expertise was developed during undergraduate education [28]. She found, however, that there were limited differences in CLASS scores (as measured by percentage of expert-like answers) between first year introductory physics students and senior students; similarly, there was no significant difference between senior students and graduate students. (The analysis is complicated, since graduate students were appreciably different from 1<sup>st</sup> year students) Moreover, a longitudinal aspect to the study found that student attitudes did not appreciably change during the first three years of undergraduate instruction. She did find significant differences between (non-major) engineering students studying physics and between the physics majors, with the former having less expert-like attitudes.

Here, we use CLASS for two purposes: to track any possible growth or backslide in student attitudes towards physics over the length of the 1-Series courses, and to establish where 1-Series students lie in aggregate on the Gire's continuum between engineering students and physics graduate students. 1-Series students are not majors, so there is no reason to expect their interest and focus to be on physics; still, by examining individual survey items and comparing to Gire's findings we can gain a more nuanced understanding of where they are more expert-like.

An initial possible objection to considering student performance on these surveys as meaningfully representing their own viewpoints is the suggestion that students are just answering what their instructors want to hear. Although students have certainly been known to contrast the physicists' models with "real life", that doesn't appear to be what is going on in these attitudes surveys. The students' own (non-expert) scores imply otherwise; students aiming to please but doing "poorly" would need to be completely misunderstanding the expert (instructor-preferred) attitudes to such an extent that their non-expert designation would be appropriate. More specifically, the CLASS team performed a clever experiment by asking students to respond to the survey as themselves initially, and then respond as they thought their professor would [24]. The students responding as themselves scored as typical non-expert students, and their "professor" answers successfully emulated expert attitudes, even before instruction. These results suggest strongly both that typical students understand what physics expert attitudes look like, and that they reject those attitudes when answering the survey items for themselves. (Of course, a portion of the "expert attitudes" students could be simply aiming to please, and a portion of the "non-expert attitudes" students could be experts maliciously answering incorrectly, but these are unlikely to be the primary factors)

## **Summer study (full CLASS)**

Although our initial lab curriculum changes were tested in Spring and Summer 2012 for 1BL, beginning in Fall 2012 we implemented them for a full sequence of 1-Series instruction (1AL, 1BL, and 1CL). Student learning was assessed via pre- and post-instruction surveys handed out in class during each quarter. Although most of the assessment was related to measurement and uncertainty (befitting the focus of the new curriculum; see Chapter 4 for more details), we also asked CLASS questions on these surveys. The CLASS allowed us to establish a baseline of physics attitudes for the 1-Series students and to see if there were any aggregate changes in those attitudes over the course of instruction.

No lab survey included the entire suite of CLASS questions. Hand-written surveys were preferable to online surveys so that physical markings could be tracked on the forms for the measurement questions. A brief (~10-15 minute) portion of lab time was apportioned for surveys during both the second and ninth lab classes. The second lab was chosen as the pre-instruction assessment because the first lab class already contained plenty of non-laboratory tasks like syllabus reading, etc. Because of these constraints, and the inclusion of the uncertainty-based questions, only eleven of the forty-two CLASS items were included in the survey provided to the students. The particular items chosen can be seen in Appendix IV, in the Fall, Winter, and Spring surveys; the items center on student confidence in physics calculation procedures (like the new uncertainty calculation method) and in tying sense-making to those types of physics problems.

Although the individual CLASS questions used on the lab survey could still be compared to the same questions for other groups (engineers and physicists) via Gire's full CLASS survey, it

does still leave the overall CLASS score of the lab students in doubt. Is there enough information in the selected 11 CLASS questions to infer the performance on the entire survey? Can the full CLASS score be inferred from their performance of the measured subset?

To investigate this question, the full CLASS survey was given to 1C lecture students during Summer 2014. Because these students were not given the lab survey, the full CLASS was provided online via the TED course management system. In addition to providing a reasonable comparison in terms of student population to the 1-Series lab students taking the lab surveys, it allowed us to test how well the selected questions represented the entire survey.

### **Evaluating CLASS surveys**

The standard metric for CLASS-based attitude surveys has been the percentage of favorable (expert-like) responses on the Likert scale. The two types of agreement and disagreement categories were combined together to form “Agree”, “Disagree”, and the remaining “Neither” bins. Likert scales are ordinal scales, so differentiating between strong and ordinary agreement is not consistent amongst subjects. Studies have shown that this process is preferable to simply using a three-point Likert scale, because the latter leads to some moderate-feeling respondents to choose a neutral option on the scale [24].

We determined relative levels of expertise on the CLASS by examining the number or percentage of survey questions to which students provided expert-like answers. Of the 42 CLASS items, only 36 have answers with an expert consensus; the remaining six lack consensus or are questions ensuring that the respondent is reading the exam in detail (the expert consensus, where



it exists, is provided in the documentation with the CLASS survey itself, having been originally derived from sixteen professors experienced and interested in teaching physics). The percentage of students that answer in an expert-like way is recorded for each item.

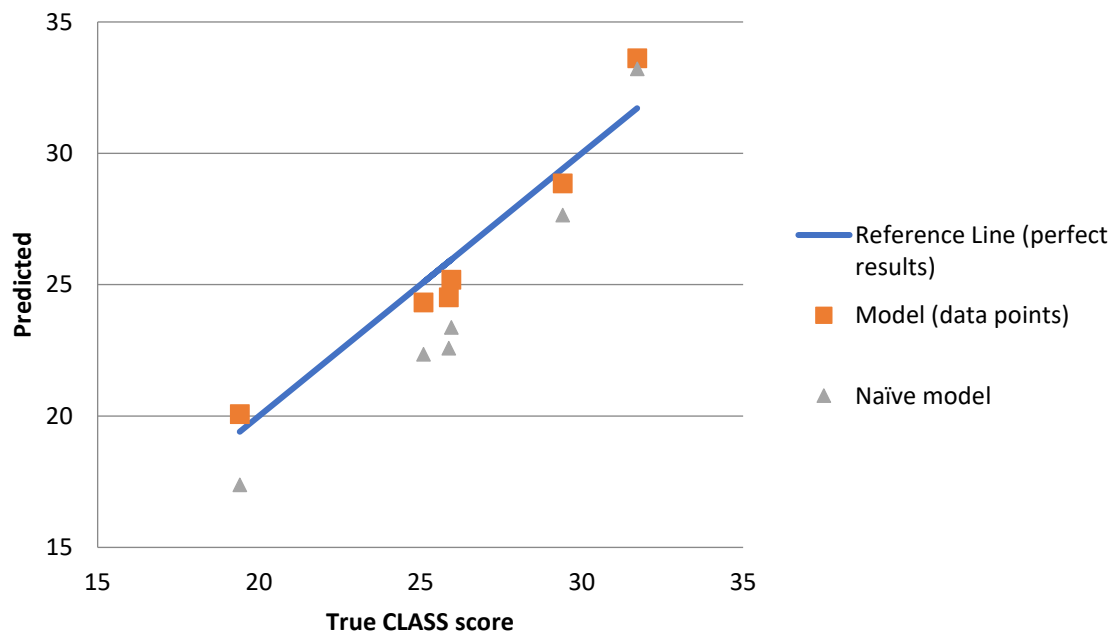
### **1. Modeling the full CLASS from our survey subset**

For the full CLASS, the summer class answered around 50% of the 36 items as experts. Students answered the 11-question subset at 42%. The correlation between these two measures was of course high, ( $r=0.84$ ), as one was a significant subset of the earlier. We determined a simple linear model with the number of expert-like responses in the subset as the independent variable, and the full CLASS as the dependent variable. (As a linear model, the correlation between these variables was also  $r=0.84$ ) This model, used in combination with the 11-question subset of the CLASS used in our uncertainty survey, allowed us to calculate the likely full CLASS results for the lab students from Fall 2012-Spring 2013. To be confident in these results, we needed to ensure that the linear model was also predictive for a different CLASS data set.

In order to check that the model applied to a completely different cohort of students, we decided to apply it to a different set of students who had taken the full CLASS. In Gire's longitudinal study of CLASS results for physics majors, engineers and graduate students, she published the CLASS expert percentages for each individual Likert item for each of her six subgroups (engineers, four years of physics students, and graduate students) [28]. For each of these groups, we tallied both the 11-question subset and the full CLASS, and used our model to predict the full CLASS results from 11-question subset. Here we calculate the CLASS score: the number of questions answered expert-like for within a study. A result of 13 on the full survey, for example,

means the student answered 13 of the 36 questions with the expert consensus answer. These scores are then averaged to determine an aggregate CLASS score.

**Graph 3-1: Graph of model-predicted CLASS score and the actual CLASS scores, based on Gire's [28] survey groups.**



In Graph 3-1, the total number of CLASS questions answered in an expert-like fashion are plotted against the model's prediction of same. The blue reference line shows for clarity if the predicted and true CLASS score were identical. Each of the squares is one of Gire's six cohorts' (the graduate students being the most expert-like in the upper right, and the engineers the least expert-like in the bottom left); the closeness of the squares to the blue line indicates the closeness of fit. For reference, the model is also compared to a naïve model using triangles, which instead

simply corrects for the total number of questions (e.g. multiplies the 11-question subset CLASS score by 36/11 to predict the full score).

**Table 3-1: Sum of squares deviation from prediction for both models.**

<b>Cohort</b>	<b>Linear model</b>	<b>"Naïve" model</b>
Gire_eng	0.44	4.10
Gire_1st	0.58	6.70
Gire_2nd	1.86	10.90
Gire_3rd	0.59	7.50
Gire_4th	0.30	3.09
Gire_grad	3.63	2.26

The sum of squares differences (distance from the line) between the two models is shown in Table 3-1. The linear model often predicts the final CLASS score to within a single item. The model only underperforms the naïve model with the graduate students, who answered on average 10.2 of the 11 questions as experts. It is not unexpected for a linear model based on a percentage to be less accurate when percentages eclipse 90%; moreover, this is the smallest of these cohorts, with only seven respondents, so the standard error in the data itself is significant.

In aggregate, the Gire cohorts strongly confirm the value of the model; indeed, the reference line even appears a plausible fit to the Gire data points, even though they were not used to determine the model fit.

There is nothing special in the specific 11 questions we used that makes our subset especially predictive of the full survey. Random 11-question selections also typically predicted overall CLASS performance as well as the subset used.

Model predictions of the expert percentages of full CLASS were as follows: 1AL Pre/Post: 56%/53%; 1BL Pre/Post: 51%/52%; 1CL Pre/Post: 57%/57%. There were no statistically significant changes within a single class. These are in-line with Gire's results for engineers of 54%.

## **2. Contextualizing 1-Series students in Gire's framework.**

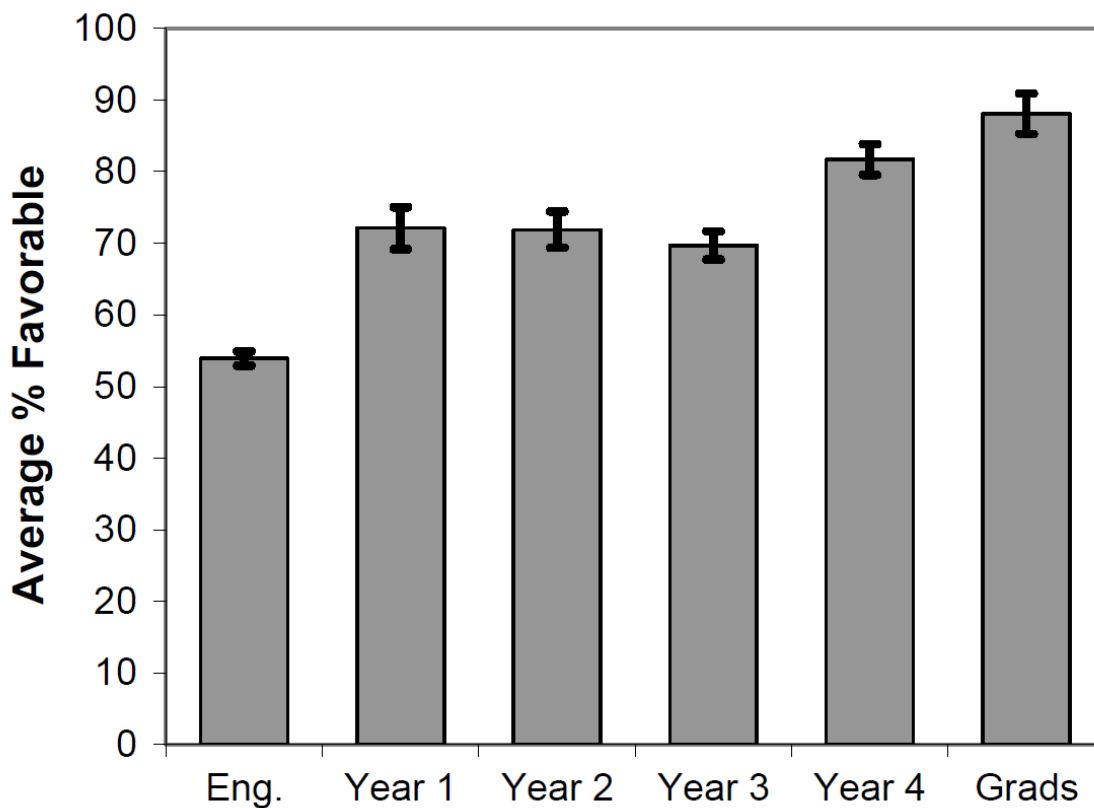
The full CLASS results gathered in the summer can be used to compare the physics attitudes of 1-Series students with those of the other physics students at UCSD: engineers, physics majors, and graduate students.

We rely on Gire's CLASS results to help contextualize the physics attitudes of students in the 1-Series lab classes for a number of reasons. First, her study examines several different cohorts (engineers, each of four year physics majors, and graduate students) that all share the same broader context as our students: they all took classes in physics at UCSD. Second, while 1-series students were expected to have fewer expert-like attitudes than engineers (who were the least expert-like of Gire's classes), the ability to compare directly to first-year physics majors allowed the examination three different groups with similar collegiate experience. Third, detailed graduate student CLASS results aid in the discussion of how appropriate the goals of physics attitudes expertise truly are for non-major students. Finally, Gire's thesis includes an item by item

breakdown for each cohort, allowing us to examine (in a qualitative way) the unusual details of our cohort aside from just relying on aggregate CLASS scores [28].

Each student's response on each CLASS item is compared to the expert consensus, yielding a percentage of expert-aligned attitudes for each student (CLASS items with no expert consensus are not tabulated). These percentages are further aggregated across the entire class, yielding an overall percentage for each measured class. 100% would be perfect agreement with Boulder's original population of 16 physics professors, which set the baseline for physics expertise.

**Graph 3-2: Gire's CLASS results reported for each cohort of physics students (from [28]).**



Overall, the CLASS results were in line with those of the engineers: 51.8% of Summer 1CL survey items were answered as experts, compared to 53.9% of engineers' items and . 72.1 % of first year physics majors' items. (In the modeled results for 1AL to 1CL based on the 11-question subset, the expert percentage ranged from 50.9% to 57.1%) Of the seven CLASS categories (see Table 3-2 for details), engineers scored higher in Real World Connection, Personal Interest, Conceptual Connections, Applied Conceptual Understanding, and Problem Solving Sophistication. The summer 1CL students scored higher in Problem Solving Confidence and Sense-Making/Effort categories. These results are on the edge of significance, as the standard error of the 1CL results is 5%. In all cases, the 1CL students were significantly less expert-like compared to the first year physics students. (Because Gire's data was tabulated during the last two weeks of instruction, all previous and subsequent 1-Series comparisons are with the post-instruction survey data)

**Table 3-2: Comparison of expert fraction in CLASS categories between 1CL students and the engineering and first year physics major students studied by Gire [28]. (Std error of 0.05 for 1CL results)**

Student Group	Real World	Personal Interest	Conc. Connect.	App. Con. Underst.	Prob. Solv. Soph.	Prob. Solv. Conf.	Sense-Making
1CL	0.55	0.45	0.44	0.34	0.34	0.57	0.64
Eng.	0.57	0.50	0.49	0.40	0.40	0.53	0.59
1 <sup>st</sup> year	0.79	0.81	0.69	0.64	0.71	0.83	0.76

### 3. Course Context

Although the engineers did perform slightly better on the CLASS, it is not by much. This may come as a surprise, considering the engineering sequence at UCSD is significantly more rigorous and time-consuming than the 1-Series courses. Engineers at UCSD have five full quarters of physics instruction, as opposed to only three for 1-Series; calculus is also more heavily integrated into their coursework. Moreover, their lab sequence includes three hours of laboratory and one hour of lecture, which leads to a more rigorous treatment of error and uncertainty, including full propagation of errors and basic statistical tests, like t-tests.

These differences in mathematical sophistication and difficulty, however, don't have much of an impact on their physics CLASS scores. Even CLASS categories like Problem Solving Sophistication are not related to the actual difficulty of problem solving, but in how the student approaches difficulties in that process. For example, both 1CL students and engineers scored low on CLASS item 5: *"After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic."* The engineers are certainly solving more difficult mathematical problems, but that doesn't mean that have more expert-like approaches to solving problems.

The major commonality of the two groups of students is that they are not majoring in physics. Since the CLASS in part is measuring motivation, interest, and conceptual focus, non-majors are less likely to engage this way with required classes that they don't think of as central to their own major. For some students, physics classes represent the required classes that they must trudge through simply because they are prerequisites for their major classes, which are more likely to correlate with the actual interests of the students.

#### 4. CLASS individual item results

Unlike Gire's engineering class, the lecture portion of the summer 1CL class was taught non-traditionally. The class consisted of 10-15 minute bursts of paired groupwork on worksheets, interspersed with traditional lecture portions. While not truly a flipped class (since pre-reading and pre-lecture preparation were minimal), it was much more interactive than the traditional lecture that the engineers had experienced. Moreover, a significant fraction of the class had also taken a similarly structured class of 1BL during the previous spring quarter.

Some of the differences between the 1CL and engineers make sense in the context of an interactive, groupwork-based class versus that of a traditional lecture. For example, consider CLASS item 19 (*"To understand physics I discuss it with friends and other students"*), where the expert response is agreement. This item is one where engineers and even physics majors are only mildly expert like: engineers and 1<sup>st</sup> year students both were at 46% expert, and this rose to only 69% for senior students. Even incoming 1CL students answered this at 74% expert, with post-instruction at 80%, less than only the graduate students at 86%. Of course, this daily student interaction and group work is the foundation of what makes the flipped class effective; moreover, their lab time is also entirely group work. Interestingly enough, this same attitude of working with others is seen as a non-expert idea in the subsequent CLASS item 20: *"I do not spend more than five minutes stuck on a physics problem before giving up or seeking help from someone else"*, where the expert consensus is disagreement. This item is fascinating, as it equates "giving up" with "seeking help" and derisively includes "five minutes" implying to respondents that harder work is required for success in physics. From another angle, however, "seeking help" could be read as seeking partners to work with, which is much more contextually appropriate in a flipped class



compared to the traditional lectures the engineers and majors found themselves in. Here, 63% of engineers and 85% of first-year majors disagree with this statement, along with 68% of incoming 1CL students (and 100% of graduate students!). But after flipped instruction, only 51% of 1CL students disagree with the proposition; which makes sense considering their in-class work experiences. Spending five whole minutes on a single problem without asking for a peer's help would be actively counterproductive in the class frame, especially when the typical group work in class occurs in bursts of 10-15 minutes. Over and over, students were encouraged to work and discuss with others, since learning is a collaborative endeavor; as the instructor of said class, I view this drop in "expertise" as success in the flipped context. This was the largest shift in the non-expert direction of any of the CLASS items.

Despite lacking expertise in their problem solving compared to the engineers, these 1CL students had significantly stronger performance on some items relating to general intellectual curiosity; some of these items fell into the Effort or Personal Interest categories, and some did not. Summer 1CL students were more expert like on items 11 (*I am not satisfied until I understand why something works the way it does*; 74% to 59%) and 15 (*If I get stuck on a physics problem on my first try, I usually try to figure out a different way that works*; 71% to 60%), while improving even to a physics 1<sup>st</sup> year level after instruction on item 13 (*I do not expect physics equations to help my understanding of the ideas; they are just for doing calculations*; 1CL students grew from 48% to 68% compared to first year majors' 63%). The most noteworthy shift in student thinking as a result of the class was a shift from 65% expert to 88% expert on CLASS item 2 (*When I am solving a physics problem, I try to decide what would be a reasonable value for the answer*), eclipsing even third year physics majors.

Although there were some interesting changes in the results of individual items, the CLASS score in aggregate remained fixed. There was a slight (non-significant) uptick in average expertise from 50.5% to 51.9%; similarly, there were no significant changes in any of the CLASS subcategories.

### **Longitudinal study (1AL-1CL)**

Surveys including the 11-question CLASS subset were collected in Fall, Winter and Spring quarters following the entire on-sequence year of 1-Series physics courses from 1AL to 1CL. Surveys were collected during week 2 and 8 to approximate a pre- and post-instruction attitudes measurement for each of the three courses; these anonymous surveys were aggregated and students were not tracked. While the goal was to determine if each class had an effect on aggregate physics attitudes, even incoming physics students may already have had differing attitudes towards the subject. Although 1-Series physics is intended to be the first introductory physics course for biology majors, some students may already have had previous experience in physics in high school. Before proceeding onward to determine the effects of UCSD's own physics curriculum on the attitudes of the students, it is important to establish if those students perceived the subject differently from those taking physics for the first time.

## **Students with previous physics experience**

CLASS item scores were compared for students in 1AL who had previously taken physics compared to those who had not. Since 1AL is the first class in the sequence, students self-reporting previous physics experience might have taken a class in high school or, more rarely, have failed it at the college level beforehand. There was a roughly even split (53%-47%) between students who had taken physics before starting the UCSD sequence and those that did not. (For ease of reference, students who had previously taken physics will be called “HS physics students”, and those who had not will be called “non-HS physics students”.) Because the CLASS deals in part with problem solving confidence and personal interest, students who have previously taken a physics class may grade more on the expert side. This may not be because the high school class itself initiated that change, but it may reflect an underlying predisposition to take the class in the first place, which might be reflected on the CLASS results. Indeed, Perkins et al. found that self-reported interest in physics correlated with CLASS results for a class of engineers; it is not unreasonable to think that students with more interest in physics in high school would have chosen to take that class more frequently [29].

Aggregating all eleven of our selected CLASS items together, there was a suggestive but not significant difference between these two groups ( $p=0.07$ ). HS physics students answered as experts 50% of the time, as opposed to only 39% for the non-HS students. This was suggestive of a difference, so we further compared individual item scores in the survey.

Since making eleven separate comparisons with a significance level of  $\alpha = 0.05$  will usually lead to some significant results by pure chance, the Holm-Bonferroni method was used to limit the likelihood of a Type I error. After calculating the p-value for each comparison, they are

sorted in ascending order and divided by their Bonferroni correction index, which is based on the number of comparisons. In this case, with eleven comparisons, the smallest p-value is compared to a significance level of  $\alpha = 0.05 / 11 = 0.0045$ . If the null hypothesis is rejected, then the next smallest p-value is tested against a slightly higher significance level of  $\alpha = 0.05/10 = 0.005$ . This continues until a null hypothesis is not rejected; all subsequent (larger) p-values would be not rejected.

For the case of comparing our 11 CLASS items between students who had taken physics before 1-Series and those who had not, four of the eleven of the item comparisons showed significant differences using the Holm-Bonferroni method; these items are shown in Table 3-3. In each case, the students with previous physics experience answered in a more expert-like way.

Although the Holm-Bonferroni test evaluates the p-values for false positives, it does not evaluate the effect size, which is how large the (significant) difference between two items are, compared to their variance. (While significance tests can determine whether two populations are different, they do not evaluate whether the amount of difference is important or large) For this, we use Cohen's h [30]. First, each proportion's arcsine transformation is calculated:

$$\varphi = 2 \arcsin \sqrt{p} \quad (1)$$

Cohen's h is defined as the difference between the two arcsine transformations of the compared proportions:

$$h = \varphi_1 - \varphi_2 \quad (2)$$

Cohen's rule of thumb for effect sizes was that h-values of 0.2 are small, 0.5 are medium and 0.8 are large. Each of the significant CLASS comparisons had medium effect sizes by this

loose metric; this makes sense, since smaller effect sizes would be less likely to be detected via the Holm-Bonferroni method given the number of students being compared ( $N_1 = 87$ ,  $N_2 = 97$ ).

**Table 3-3: Individual CLASS item differences between students who had taken physics before 1-Series and those that did not (pre-instruction). Fractions listed are agreement with expert consensus. Expert consensus is agreement for item 34 and disagreement for the rest.**

CLASS Item	Prev phys	No phys	p-val	Cohen's h
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.38	0.16	<b>0.0004</b>	0.51
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.32	0.14	<b>0.002</b>	0.44
34. I can usually figure out a way to solve physics problems.	0.52	0.26	<b>0.0002</b>	0.52
35. The subject of physics has little relation to what I experience in the real world.	0.7	0.48	<b>0.001</b>	0.46

Items 1, 5, and 34 relate to problem solving confidence and methodology in physics; in each case, the students who have taken physics before have significantly more confidence. But is it the case that the students that have taken previous students have lots of confidence, or is it instead that the students who have never taken a class before lack it? We can gauge these possibilities by comparing the level of expertise to Gire's engineers; these students agreed with the experts at a rate of 0.42, 0.3, 0.45, and 0.63 for CLASS items 1, 5, 34, and 35, respectively [28]. In each case, the confidence of the HS physics students is comparable to that of the engineers; moreover for item 35 they have a similar opinion about the relationship of physics to the real world. By contrast,

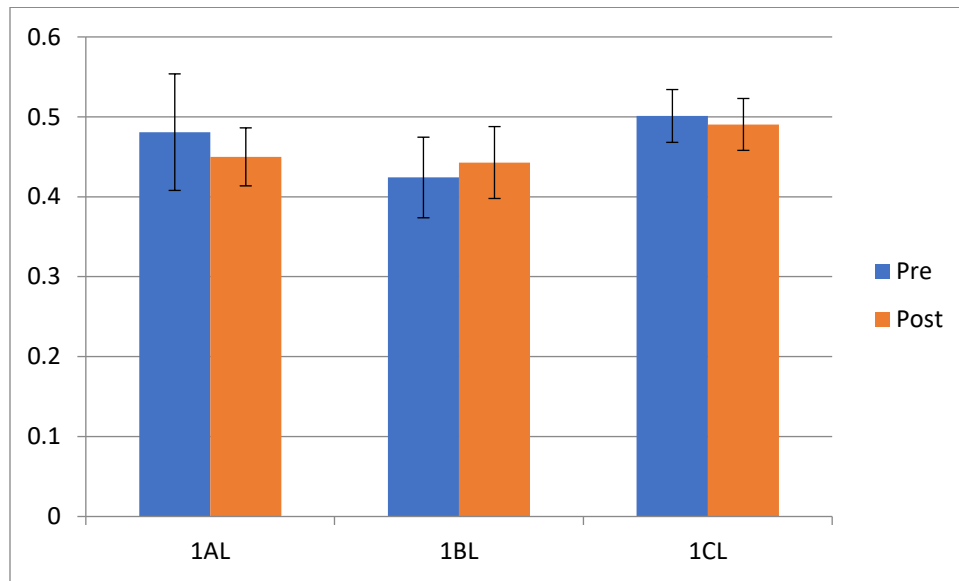
the confidence of the students that had not taken high school physics is in fact quite low. There is no significant difference in these CLASS items before or after instruction for either set of students, so these results don't simply reflect a lack of confidence due to not having been in the class before.

Most high schools [31] in the country do offer physics classes. Some students may have chosen to avoid physics classes in high school rather than being unable to take them. That these students also have significantly less expert-like scores on the problem solving confidence items may point to a lack of mathematical confidence. (They rate as expert-like as the rest of the 1AL students on conceptual connection items such as “*In physics, it is important for me to make sense out of formulas before I can use them correctly*” and “*When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.*”) For the HS students, Perkins et al., found that among students who became more interested in physics after an introductory class, they commonly attributed that to physics' connection to the real world or because they believed it would be useful [29]. One can imagine these causes may have also been at work for the high school physics students who would later take 1AL at UCSD, thus explaining their higher subsequent scores on the CLASS.

### **Attitudes changes within individual 1-Series classes.**

Overall, there were not any significant changes in any of the measured CLASS items (in aggregate, individually, or projecting to the full survey) between pre- and post-instruction surveys for 1AL, 1BL, and 1CL taken individually. (This does not mean there were no measurable changes in any survey performance, only in the CLASS portion. See Chapter 4 for more details on the uncertainty measuring portion of the surveys)

**Graph 3-3: CLASS 11-question subset results for each class in the sequence. The fraction of expert-like responses is shown; error bars are standard errors of the mean.**



One survey item did have a nearly significant change ( $p\text{-value} = 0.008$ ), which was item 35 (*The subject of physics has little relation to what I experience in the real world*). For 1AL students, the expert response to this dropped from 79% before instruction to 60% post-instruction. This change would have a large effect size ( $h=0.4$ ); but is likely not significant because pre-instruction had a low survey return rate ( $N=46$ , compared to  $N=179$  post-instruction). This was one of the survey items that was significantly different between 1AL students that had taken physics (expert 70%) vs those that had not (expert 48%). In the 1CL survey, we continued to ask students if they had taken 1AL during that same year; those students answered as experts 61% of the time, as opposed to those 1CL students that had taken 1AL at an earlier time (expert 46%).

This difference was not significant because of the Holm-Bonferroni test, but also had a low p-value of 0.006.

If these differences are suggestive, what could they mean? Item 35 seems very dependent on the content and presentation of the class: a highly abstract version of a mechanics class could feel not very grounded in the real world compared to a more demo-heavy version, for example. Still, these differences need not be attributed to, for example, the way that physics was presented in high school for those that took it there. After all, those students may have taken that physics class because they believed the science to be connected to the real world compared to those that demurred. Either way, it appears that the remnants of this split are still there six months later when the same students are taking 1CL.

One should be careful, however, about comparing results from that Spring 1CL class to previous classes. The students in that group have taken three straight quarters of physics and so potentially there is a survivorship bias when comparing them to students who have delayed their third quarter of physics. This sort of class composition change can be seen most clearly when comparing students in that 1CL class who had just finished taking 1BL in the Winter vs those that had taken it previously at some other time.



### Transition between Winter 1BL and Spring 1CL

Even though the 1BL and 1CL classes had similar physics attitudes overall, three CLASS items were significantly differently ( $p < 0.005$  for each) even by the standards of the Holm-Bonferroni test. Again, these were CLASS items 1, 5, and 34 (see Table 3-4 for the item statements) which are related to problem solving.

**Table 3-4: Individual CLASS item differences between students at the end of 1BL and at the beginning of 1CL. Fractions listed are agreement with expert consensus. Expert consensus is agreement for item 34 and disagreement for the rest.**

CLASS Item	Post-Instr. 1BL.	Pre-Instr. 1CL	p-val	Cohen's h
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.30	0.43	<b>0.003</b>	0.27
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.24	0.35	<b>0.002</b>	0.29
34. I can usually figure out a way to solve physics problems.	0.44	0.59	<b>0.001</b>	0.30

In each case, the 1CL class answered in a more expert-like way. There is a gap of less than a week between the end of Winter quarter and the beginning of Spring quarter, so these changes are not likely to represent any changes in student views of physics. Instead, this is likely due to changes in class composition between the quarters. Even though most of the students have been retained (79% of the 1CL students had taken the Winter 1BL class and 71% had also been in the

Fall 1AL class), there is still an infusion of students whose attitudes had not been previously measured.

But those two groups are not distinguishable in the 1CL survey data: there are no significant differences on those CLASS item scores between 1CL students that took 1BL immediately beforehand in the Winter and those that had taken the class in a previous quarter. What leads to the higher 1CL attitudes scores on these items, then? One hypothesis is a population of relatively non-expert attitudes students postponed the 1C class after having finished 1BL in the Winter. There is no direct survey data on their attitudes as a group, but 1B (electricity and magnetism) is often a taxing class for students and it is easy to imagine some degree of burnout following that class. Could their absence be enough to raise the expertise scores that much on the 1CL results?

These students may consist of a significant fraction of the 1AL students that had previously taken no physics class in high school. These students had very low expert attitudes in three of the same CLASS items, as seen in Table 3-3; assuming that those attitudes did not change in between classes, it is likely most of those students are still present in 1BL (Around 50% of 1AL students hadn't taken a physics class before; 86% of 1BL students reported taking 1AL in that previous fall). However, there was also a significant difference in for CLASS item #35 (*The subject of physics has little relation to what I experience in the real world*): only 48% of non-HS students disagreed with that statement, as compared with 70% who had taken high school physics. Why is this difference not also seen in the 1BL-1CL transition data, if the non-high school physics students are the cause of those differences? After all, the 1CL (pre-instruction) students disagreed with that statement at 51%, compared to 54% for the 1BL students; this is a non-significant difference, and is even in the wrong direction!

Further insight comes when reexamining the 1CL students that had just taken 1BL in the Winter versus those that had taken it in a previous quarter. Although there are no significant differences in CLASS items between those groups, item #35 comes close. Students who had just taken 1BL disagreed with that item at 61%, compared to only 43% who took 1BL at a different time. The p-value of this difference is 0.005, just missing the Holm-Bonferroni threshold of 0.0045 for eleven comparisons; still, the effect size is worth noting at  $h=0.35$ . Similar results can be found in a comparison of 1CL students that had taken the entire series since the Fall;  $p=0.006$ ,  $h=0.31$ . This is the only comparison even close to significant between these subgroups of 1CL.

One possible mechanism for the increase in those CLASS item scores between 1BL and 1CL is as follows. A large fraction of the students in 1BL that had not taken physics in high school (and thus scored lower on those CLASS items) chose not to take 1CL on sequence in the Spring. Consequently, CLASS scores in aggregate rose on those items. Although that subgroup also scored lower on another CLASS item (#35), when they left the class they were replaced with 1CL students (that had taken 1BL in an earlier quarter) who had also scored low on that item, so the aggregate 1CL score on that item did not change in comparison to the 1BL results.

Of course, students that decide to take classes off-sequence do have to complete the classes they defer at some later point. But because delaying classes can affect their degree plan with their major, often these classes can be postponed until the summer. Because our CLASS subset validation study (see above) occurred in a summer 1CL class, differences in attitudes might serve as tracers of that kind of student movement.

## Comparing Spring 1CL to Summer 1CL

While these two classes have the same content knowledge and draw from the same overall pool of students, through circumstances or choice they took the class at a different time. The suggested degree path (see Chapter 5 for more details) for biology has students take 1AL-1BL-1CL back to back to back in the same academic year, often before their biology coursework has begun. The first two years of a biology degree are mostly out of major degree requirements: math, physics and chemistry, along with labs for the latter two. If a student is feeling overwhelmed or discouraged at this intense sequence of classes, some class needs to be deferred until after the biology classes begin. The only possibility is physics: chemistry classes are often prerequisites to upper division biology classes, but physics 1-Series is only a degree requirement. Consequently, it makes sense in certain situations to postpone those classes; summer students are more likely to be juniors or seniors than on-sequence students.

Overall, pre-instruction CLASS 11-item scores were slightly lower (in a non-significant way) for the Summer students compared to the Spring students. Still, there were several individual items that showed significant differences; these can be seen in Table 3-5, along with two others with p-values small enough to discuss.

In each case, the summer students were less expert-like. Again, CLASS items 1, 5, and 34 appear, their low scores reminiscent of the confidence of the students that did not take physics in high school (as seen in Table 3-4). It is hard to examine these numbers without writing narratives. For example, CLASS item #34: (*I can usually figure out a way to solve physics problems*) 71% of the students in the Spring had just completed two 1-Series classes and had committed to the third. These remaining students would almost have to have some confidence in their abilities vis-à-vis

physics problems; if not, they probably would have pushed off that last quarter class! Anecdotally, some of the Summer 1CL students had dreaded taking the class, pushing it off as far as possible. And since physics requires some continuity of skills related to subject matter, some lack of confidence that one has forgotten what one has previously learned is natural.

**Table 3-5: Comparison of % expert responses to CLASS items in Spring and Summer 1CL classes. Fractions listed are agreement with expert consensus. P-values in bold are significant using the Holm-Bonferroni test.**

CLASS Item	Spring 1CL	Sum. 1CL	p-val	Cohen's h
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.43	0.30	0.016	0.27
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.35	0.18	<b>0.0016</b>	0.39
6. Knowledge in physics consists of many disconnected topics.	0.64	0.41	<b>&lt; 0.0001</b>	0.47
12. I cannot learn physics if the teacher does not explain things well in class.	0.20	0.05	<b>0.0005</b>	0.48
34. I can usually figure out a way to solve physics problems.	0.59	0.36	<b>0.0001</b>	0.46
42. When studying physics, I relate the important information to what I already know rather than just memorizing the way it is presented.	0.64	0.77	0.014	0.28

CLASS item #12 (*I cannot learn physics if the teacher does not explain things well in class*) was universally low for 1CL students: at most 20% of students disagreed with this statement and felt they had enough confidence and motivation to learn the subject on their own. (Gire also found a similar attitude even for physics majors; only graduate students disagreed with this statement at more than 50%.) Of course, this could also represent the discouraging effect of poor teaching,

where a student perceives their difficulties as being due to the teaching. Perhaps in one's major, one would be more willing to push through, since the topic knowledge may be necessary for their future; of course, this is less likely to be the case in a resented course requirement like physics for biologists. The summer students were well below even this low expert rate; only 5% of the summer students disagreed with the statement! <sup>6</sup>

It is certainly possible that some of these relatively low scores for the summer students may be attributable to previous physics classes; the typical effect of a traditional lecture is a reduction in CLASS expertise and most physics classes at UCSD are traditional lectures. Additionally, Perkins found that the most common reason for students to report decreased interest in physics after instruction was due to specific aspects of instruction, such as course difficulty and teaching style [29]. That self-reported decrease in interest was correlated to students' CLASS performance, so negative course experiences may cause students to answer the CLASS in a less expert-like way. Since there are multiple different lecture sections operating simultaneously (and our data were taken in the mixed lab class setting), there is no straightforward way to proceed with a class by class analysis. Conversely, there is some evidence that the differences in attitudes between students who took physics in high school and those that have not are still perceptible in the attitudes of different classes; but that distinction cannot explain items like #6 and #12 which were not noticeably different between those two groups. More study is required to see if these differences are attributable to individual class effects, differences in the incoming student body, or some other more subtle distinction.

---

<sup>6</sup> Ironically, this very low result may be partially attributable to the more active learning found in the partially flipped class. Even though the students are better collaborators in their own learning in such a setting, they may also attribute its efficacy to the professor rather than their own efforts.

## **Student workload**

An initial concern in the planning of the curriculum changes was that the new uncertainty tasks might add too much additional work for the students. After all, the additional techniques and reasoning were not supported by the lecture portion of the class at all; while some amount of the practice came during class, much of it would need to come from out of class time during the writing of lab conclusions. Although some effort was made to limit the scope of those conclusions (both by limiting them to a single page, and by including prompts), we may have been underestimating the amount of additional work that would be required.

Rather than simply looking at a raw measure of the number of hours students spent on the conclusion, we were interested in whether students felt the added work was too much compared to the previous lab conclusions. An additional survey item (answered via a five point Likert scale) was included: “The conclusions for the lab class in general took more time to complete than I thought they should”. This survey was provided to students in 1BL for the original pilot study, who had previously taken an unmodified version of 1AL. In the pre-instruction survey, they answered for their previous experience in 1AL; in the post-instruction survey, they answered for the new lab conclusions in 1BL.

In general, most of the students (79%) believed the workload in 1AL was already too high, which matches anecdotal discussions with students in office hours. The percentage of students agreeing with that statement, however, dropped to 69% after instruction; this change was not statistically significant ( $p=0.28$ ). (This question was asked each quarter after instruction, and did

not deviate from the 60-70% range) Since the goal in asking this question was to ensure that students were not more unsatisfied with the new workload, we did not look in to the issue further. To judge by discussions in office hours both before and after the switch, following a set prompt with a consistent structure was preferable to the sometimes vague instructions provided by their previous teaching assistants. The high absolute percentage of students feeling overworked may stem from the basic structure of the class: a lab course that meets each week for two hours and requires a lab conclusion may simply be perceived as too much work for a 2-unit class.



## **Chapter 4 : Uncertainty probes**

### **Introduction**

The course modifications began in a pilot study during Spring 2012 in 1BL (the electricity and magnetism class). These in-class instruction and conclusion changes were later migrated to the other courses in the series, starting in Fall 2012 in 1AL, and following that same cohort through Winter 2012 (1BL) and into Spring 2013 (1CL; waves, optics, etc.). Surveys were distributed twice in each quarter: once near the beginning of the course (during the second lab), and once near the end (during the eighth or ninth lab). Surveys contained a mixture of Likert-scale measurements of attitude towards physics (See Chapter 3 for more details), and experimental probes measuring student understanding and skills related to uncertainty. The surveys themselves can be found in Appendix III.

Survey tools were designed to directly evaluate how well students could estimate and use uncertainty in a series of plausible laboratory-based questions. The major categories of such inquiry included: ruler measurement, graph interpretation, and usage of spread-based reasoning to draw conclusions about data. There was no lab practicum element: relevant data were shown to the students on the printed surveys. The surveys were taken in class and were anonymous, limiting the ability to examine pre- and post-survey changes in specific students. In general, the focus was not to use a small set of repeated tools to track longitudinal changes in the whole cohort of 1-Series students, but to use several distinct methods to try to measure the breadth of student skills

concerning uncertainty. Despite this focus, some statistically significant changes in aggregate student performance were seen with some probes.

In this chapter, we first examine the pilot study in 1BL, whose survey centered on an estimation of uncertainty using a simulated ruler measurement. Next, we look at the later modifications of that task, including how students determined their uncertainty numbers. Finally, we present a variety of diverse tasks, which examined how student skills and perception of measurement uncertainty changed via different levels of priming and complexity, including how to use uncertainty to decide whether data sets agreed.

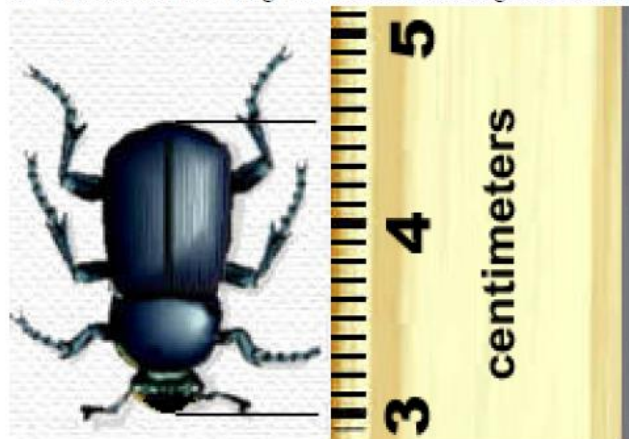
### **Pilot study on ruler measurement**

Ruler measurement activities have often been used as tracer of understanding of measurement uncertainty in the literature. In some cases, like Deardorff's study [8], the focus has been on students measuring an object with a physical ruler. These approaches measure both (i) the procedural knowledge required to actually make the measurement correctly, involving physically manipulating the measuring device and the object to be measured, and (ii) the evaluation of the uncertainty from both observations of the process and estimating any other relevant factors (such as inconsistencies in the measuring device or object or any possible systematics). The physical effects on their own are complicated – for instance, engaging in goal-driven actions with an object influences perceptions of its size [32] – so in the 1-Series surveys the interpretation has been isolated as much as possible from physical interaction with an object. Moreover, a significant portion of in-class data measurement was through digital measuring devices, like multimeters,

LoggerPro, or other computer data acquisition systems (DAQs). Focusing purely on physical ruler measurement wouldn't have traced student development in these lab tasks. Instead, objects are shown on printed surveys next to a measuring device and students were given the task of measuring both the average value and the uncertainty of the object, according to the ruler. (Printed estimation tasks emulating computer data acquisition can be found later in this Chapter; see "Determining uncertainty from a graph")

**Question #6: Reading a ruler (RR)**

6. Which of the following *best* describes the length of the beetle's body?



The beetle's body is

- a) between 0 and 2 cm long
- b) between 1 and 2 cm long
- c) between 1.5 and 1.6 cm long
- d) between 1.51 and 1.55 cm long
- e) between 1.525 and 1.535 cm long

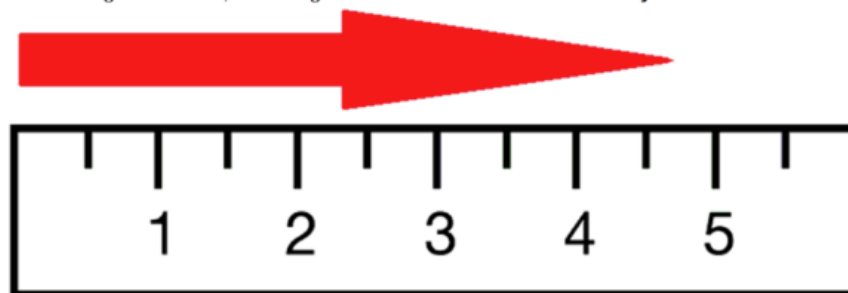
Explain your choice.

**Figure 4-1: Abbott's [12] ruler measurement task.**

This sort of limited ruler-measurement task has been explored by Abbott [12], using the probe depicted in Figure 4-1. Here, the measurement factor was almost entirely eliminated, in favor of a laser-like focus on estimating the uncertainty. Lines indicating where measurement should begin and end were provided, making the measurement less of the beetle and more of the horizontal lines themselves. The beetle was placed far away from the zero measurement point to disambiguate ruler measurements of a certain part of the beetle's body from its overall length. Moreover, by providing the ranges of possible values as multiple choice options, the details of the estimation procedure are elided.

An intermediate choice between Abbott's tool and physical measurement was chosen, which minimizes the physical parts of the measurement but leaves open enough of the measuring process to investigate students' lab approaches towards determining an uncertainty. Digital measuring devices only sometimes calculate a measurement uncertainty, so student estimation of uncertainty applies both to analog and digital devices.

7. How long is the arrow, according to the ruler? What is the *uncertainty* of that measurement?



**Figure 4-2: Pilot study ruler measurement task.**

The most significant difference with standard ruler measurement probes is the intentional lack of intermediate tick marks on the ruler between the half-tick marks. This exaggeration of the space between marks requires a careful, sense-making estimate of both the length of the arrow and its uncertainty that is not determined by the spacing of the marks alone. In real-world situations where precision of reading on these sparse scales is paramount (e.g., pilots reading instrument dials), recording the amount of uncertainty due to lighting and parallax has been well-studied [33]; these effects were minimized by having students not physically perform the ruler measurement.

In the initial pilot study, students were only asked to determine the length of the arrow and its uncertainty. The study was performed in Spring 2012 in 1BL, which was the first quarter using the new uncertainty approach towards the lab classes. While these students had previously had a quarter of physics, it did not contain a detailed or specific focus on uncertainty. Surveys were given both pre- and post-instruction.

## **1. Results**

In both the pre-instruction and post-instruction surveys, students were provided with a picture of an arrow and ruler as seen in Figure 4-2, and asked to use the ruler to determine the length of the arrow and the uncertainty of their measurement. The tick marks on the ruler are quite far apart, making them not very useful for the task of determining the uncertainty of the length of the arrow; therefore, some sort of estimation or range procedure was required to determine a reasonable uncertainty. A common rule of thumb is that the uncertainty of a ruler measurement is equal to the distance between the measuring increments (or half that distance). This rule becomes

less and less precise the further the measuring increments are apart, as it doesn't interpolate at all between the marked increments.

In the survey item shown in Figure 4-2, the right edge of the arrow is in between the 4.5 in and 5.0 in markings on the ruler. This rule leads to an uncertainty of  $\pm 0.5$  in or  $\pm 0.25$  in, depending on the version of the measuring increment rule. The result of using this methodology, then, would be  $4.75 \pm 0.5$  in or  $4.75 \pm 0.25$  in.

This method, of course, is not very appropriate for the task at hand, because the measurement increments are so far apart. Moreover, the extreme values implied by these ranges are outside the realm of reasonableness: the maximum value of 5 in or 5.25 in implied by those ranges is clearly far past where the arrow reaches <sup>7</sup>.

Skilled estimators (or skilled dial measurers), however, should be able to determine the length of the arrow to higher precision. This is often an iterative process: by noticing when the initial uncertainty estimate yields edge ranges that are too extreme, the expert can reduce that uncertainty until the upper and lower limits are within reasonable bounds. Better estimates of length and uncertainty can be found by further subdividing the ruler by marking it on the paper, and then following the same procedure. As can be seen in Table 4-1, the average expert uncertainty was less than a quarter of the 0.25 in ruler increment approach.

To determine a baseline level of expertise for this task, graduate student lab instructors (and coordinators) were asked to estimate both length and uncertainty for the same images. Three

---

<sup>7</sup> One might object to the placement of the average precisely in the middle of the interval, since the arrow appears to be closer to 4.5 in than to 5.0 in. With such large uncertainty ranges, however, any average length in this range will mean edge values outside of the measuring increments anyway.

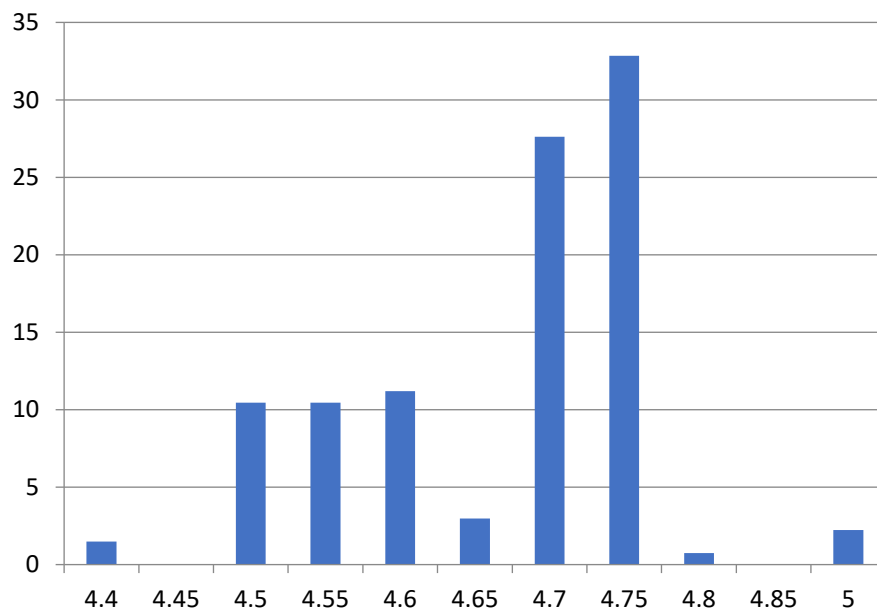
of the graduates stand out: they each choose a value of length precisely between the two marked increments, and an uncertainty that measures to those increments ( $4.75 \pm 0.25$  in). The full range corresponding to these estimates is 0.5 inches, which is just the distance between the tick marks. This is just the naïve increment measurement approach described above. The others, however, appear to be actually estimating or using range-based procedures, as they have distinct values of both length and uncertainty. (The 0.0625 uncertainty estimator divided the ruler into sixteenths) These others define our “expert” sample for this study, representing some application of expertise in determining both the length and especially the uncertainty.

The expert TAs each chose an average length less than 4.75 in, as the arrow appears closer to the 4.5 in increment than the 5 in increment. Their estimates of the uncertainty were quite small, ranging from 0.03 to 0.1 inches. Were these uncertainty estimates intended as 1-sigma results or 95% confidence intervals? Although their intent was not recorded, an indication of it can be seen by comparing the standard deviation of the mean estimates (0.026 in) to the mean of their uncertainty estimates (0.059 in). Their uncertainty estimates are around twice as large as the spread of their means, so it seems likely that these estimates are closer to 2-sigma uncertainties. Using either the 2-sigma standard deviation of the mean or the self-reported uncertainties yields a 95% confidence expert interval between around 4.63 in and 4.73 in. (But no expert estimates of length were above 4.7 inches, as befits the arrow measurement that is clearly shorter than 4.75 inches).

In aggregate, the length estimates of the students were very similar to that of the expert TAs: 4.675 in vs 4.682 in. This is a bit misleading, however, as can be seen on the histogram of student length measurements as seen in Graph 4-1. Out of the 134 students who determined a length, only 42 chose a value within the expert TA 95% confidence limits; 47 chose values too large, and 45 too small. Still, overall the students chose fairly reasonable values for the length of

the arrow. Even though 4.75 in is too high a value of the length for an expert, it is still less than 0.1 inches away from the expert length; if this wider interval were chosen, 71% of students made acceptable length determinations.

**Graph 4-1: Histogram of the percentage of students estimating arrow length pre-instruction (N=134).**



Student estimates of uncertainty were considerably less expert-like. Many of the students didn't even answer the uncertainty part of the question. Of the 139 students surveyed, 45 did not estimate the uncertainty, while still filling out the other parts of the survey. (Only two students provided a value for the uncertainty without estimating the length of the arrow.) Misreadings of the question are not the only contributor to this high number, since some of the students who did not provide a numerical answer to the question objected to it in text, saying for instance that “we



**Table 4-1: Length estimates and uncertainties for all TAs, expert TAs alone, and pre-instruction students (N=134 for length, N=94 for uncertainty).**

Teaching Assistants	Length(in)	Unc. (in)
	4.6875	0.0625
	4.7	0.05
	4.69	0.03
	4.63	0.1
	4.75	0.25
	4.7	0.05
	4.75	0.25
	4.75	0.25
Mean	4.71	0.13
Std. Dev.	0.04	0.14

**Expert TAs only  
(N = 5)**

Mean	4.6815	0.0585
Std. Dev.	0.026	0.023

**Students**

Mean	4.687	0.27
Std. Dev.	0.209	0.478

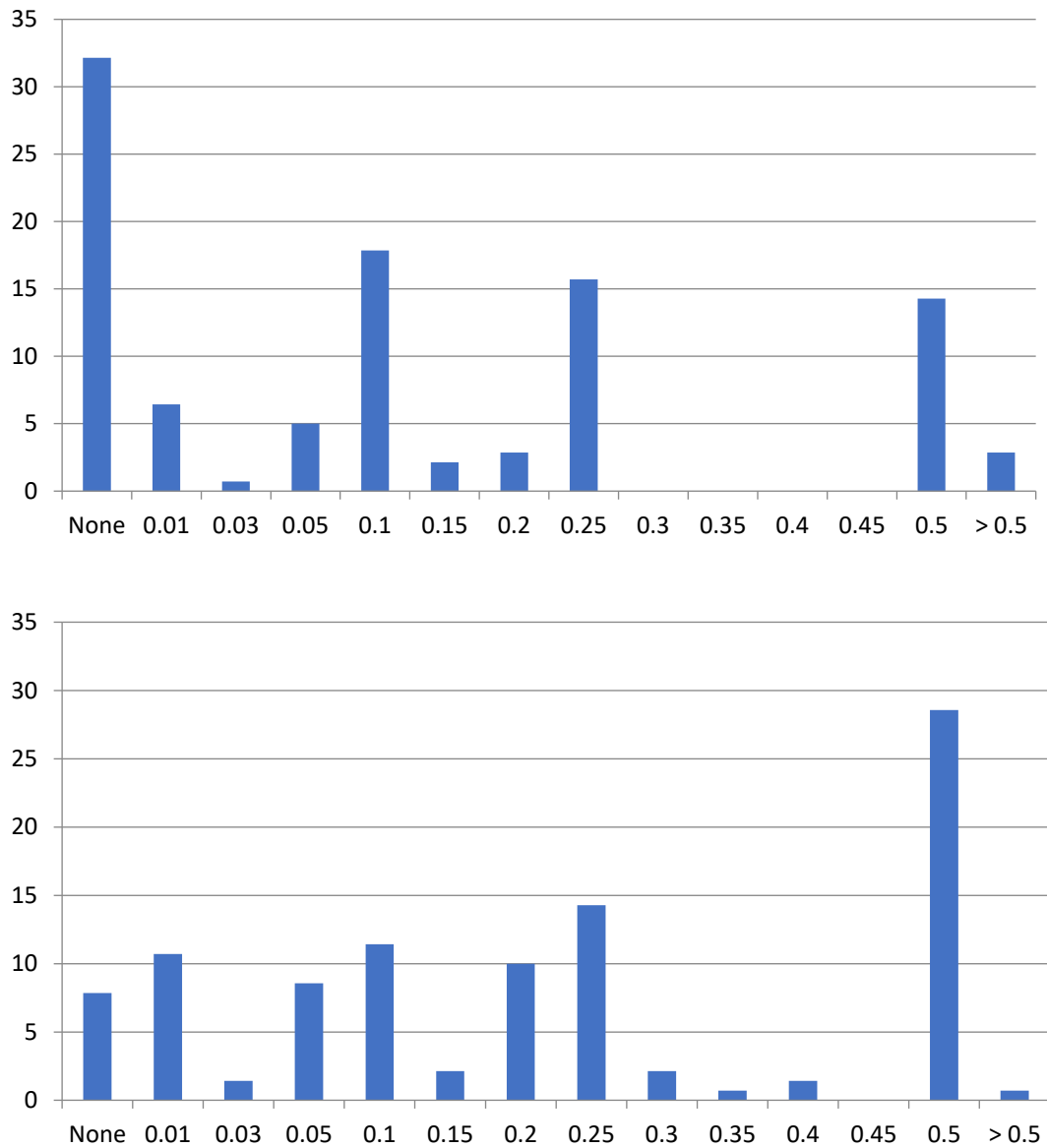
don't know where exactly the arrow starts and ends.” One student felt the length of the arrow was estimable, but not the uncertainty: “The arrow is about 4.75 inches. However, there is an

uncertainty that lies in the vagueness of measurements on the ruler. It doesn't specify what units are moved and smaller intervals." If these smaller intervals were not provided, the student was unwilling or unable to estimate the length to higher precision. On the post-instruction survey, the tendency was mitigated: only ten (out of 142) students wrote an average without a numeric uncertainty (significantly different at the  $p=0.0001$  level), and four did the converse. This may be attributable to increased familiarity of uncertainty itself, or at least of the language the class used to discuss it.

Most student uncertainties were round-numbered values. Most students chose either 0.5, 0.25, or 0.1 as their uncertainty. The large values of 0.25 in and 0.5 in may have been attributable to the smallest increment method discussed earlier. And while the 0.1 inch uncertainty may be a (slightly large) estimate of the uncertainty, it might also be the use of a different heuristic: one unit in the smallest decimal place of the uncertainty estimate. Indeed, students that chose a 0.1 in uncertainty were more likely to also have chosen a one decimal place length estimate (68% vs 43% of the other students; significant at  $p=0.02$ ).

In the post-instruction survey (where students examined a different arrow, and experts estimated a length of  $3.62 \pm 0.1$  in), the most common uncertainty estimates were still at the specific round numbered values of 0.1 in, 0.25 in, and 0.5 in. Unlike in the pre-instruction survey, where these three values were the only common choices, students after instruction were more likely to choose values in between these main values. Even if the mean student uncertainties are similar, examining these intermediate values can serve as a tracer for students using the more expert-like techniques of estimation or perhaps even range-based approaches.

**Graph 4-2: Percentage histogram of pre-instruction (above) and post-instruction (below) student determinations of the uncertainty in the length of the arrow.**



## 2. "Sure Estimators"

Still, nearly 50% of students chose an uncertainty of either 0.1, 0.25 or 0.5 inches even after instruction. While it is unclear whether students choose these common uncertainties because they were estimating or applying some other procedure that leads to the same answer (e.g., always using half of the separation between tick marks), students who select intermediate values seem to be actually estimating the uncertainty of the length of the arrow. An answer of “0.4 in” for instance, implies an improbably large range of values, but one can be confident that such an unusual value was not likely to be the result of a formal rule religiously applied. These intermediate bins, considered alone, approximate the number of students that can be reasonably supposed to be estimating – thus “sure estimators”.

In the pre-instruction survey, only 13% of students are “sure estimators”. Of course, it is likely that at least some of the students choosing 0.1 in and 0.25 in just happened to estimate those numbers for the uncertainty, instead of applying the rigid rules that are inappropriate for our ruler; these estimation numbers are likely to be lower limits, for this reason.

After instruction, the percentage of sure estimators doubled to 27%. This is not simply due to more students answering the uncertainty part of the question. Even if the non-answerers are excluded, the number of sure estimators rises from 19% to 28%, which is still a significant increase at  $p=0.01$ .

Although the students do appear to be more comfortable with determining the uncertainty (and, probably more willing to estimate uncertainty values), the average value of the uncertainty over the population remains very similar, as seen in Table 4-2. The increase in values less than 0.1 inches is counteracted by the large jump of 0.5 inch values, as can be seen in Graph 4-2. These changes were significant: the number of 0.1 values decreased ( $p=0.01$ ), and the values at 0.5

increased ( $p=0.04$ ). Similar changes and significance levels were also found for a subsequent summer study of the same surveys forms.

**Table 4-2: Uncertainty estimates of the arrow for expert TAs and students.**

	Pre-Inst. (in)	Post-Inst. (in)
Experts	0.06	0.10
Students	0.28	0.26

### **3. Comparing student distributions with expert distributions**

Even though the length and uncertainty can be examined separately, comparing the full ranges themselves represents a better summary of student skill. Consequently, our approach here is to find a metric to compare each student's range with that of the experts.

While the initial instinct was to simply calculate the overlap (or convolution) of the student ranges and the expert ranges, this led to some unintuitive results given what the data looks like. Because student uncertainties were often so large (e.g., 0.5 or 0.25 inches), their ranges would often entirely engulf the expert spread. This would result in a very small value regardless of whether the lengths were similar or different. In cases where student uncertainty was smaller, the overlap was also usually small, again giving a very small value for the result, even if qualitatively this result was better to our eyes than the large uncertainty choices. There is an asymmetry here: smaller uncertainty values should be “better” than very large uncertainty values, which was rarely

the case with an explicit overlap calculation. We<sup>8</sup> looked for another way of comparing two ranges that were sensitive to how these data sets behaved, and in-line with our qualitative perceptions of which student results were closer to the expert values.

The Kullback-Leibler (K-L) divergence is an asymmetric construct that measures the amount of information lost when approximating one distribution with another [34]. It is defined as:

$$D(p_0 \parallel p_1) \equiv \int p_1(z) \ln \frac{p_1(z)}{p_0(z)} dz \quad (1)$$

where  $p_1(z)$  and  $p_0(z)$  are the probability densities of the two distributions. (In our case, these are the student and expert distributions.)

This allows the comparison between each student's average and uncertainty (expressed as a Gaussian distribution) and the combined graduate students' values for the same, which serve as the expert distribution. This K-L divergence, then, can represent what is lost by using a student range instead of the expert range.

Because the K-L divergence is asymmetric, it cannot serve as a distance between the two distributions. In other words, the information lost by approximating the student range by the expert range is different than the information lost when doing the converse.

Johnson and Sinanović [35] have provided a way of symmetrizing the divergence in the following way:

$$\frac{1}{R(p_1, p_0)} \equiv \frac{1}{D(p_1 \parallel p_0)} + \frac{1}{D(p_0 \parallel p_1)} \quad (2)$$

---

<sup>8</sup> Joe Salamon was invaluable in finding and applying the K-L divergence to student uncertainty distributions.

where  $R(p_1, p_0)$  defines the resistor-average distance. The advantage of using a K-L divergence based measure is that this divergence is calculable in closed form for the comparison of two Gaussians. Simply summing up the two K-L divergences connecting the two distributions is possible, but it tends to overemphasize the large divergence values at the expense of the small values. Directly summing a large and small divergence value yields a sum near to the large value, which grouped otherwise dissimilar student estimates together. With the resistor distance procedure, the resistor distance sum is closer to the small value, which helps to differentiate the data a bit better.

In order to calculate the divergence, student averages and uncertainties need to be expressed as Gaussians; this means we have to decide whether student uncertainty values are better approximated as 1-sigma or 2-sigma (95% confidence) intervals. Because no statistics were taught in the class, strictly interpreting student uncertainty estimates as 1-sigma standard deviations might be misleading, as that would mean 32% of results are outside of that data range. The class procedure for determining uncertainty typically involved measuring 3-5 times, and then using half of this full range of measured values as the uncertainty for the measurement. This means that the typical uncertainty (given Gaussian distributed measurements) was usually between 1 and 1.5 standard deviations of the statistical uncertainty of those measurements. Using the student uncertainties as 1-sigma, then, is reasonable as a first approximation<sup>9</sup>.

The distribution of K-L divergence distances (“resistor distances”) is shown in Graph 4-3. What do these numbers mean and what values are “good”? Since the K-L resistor distance values represent the information lost when approximating one range by another (in this case, individual

---

<sup>9</sup>Using the uncertainty estimates as 2 sigma intervals doesn’t appreciably change the subsequent data results, except to increase the distance values between the estimates (since the effective sigma is smaller).

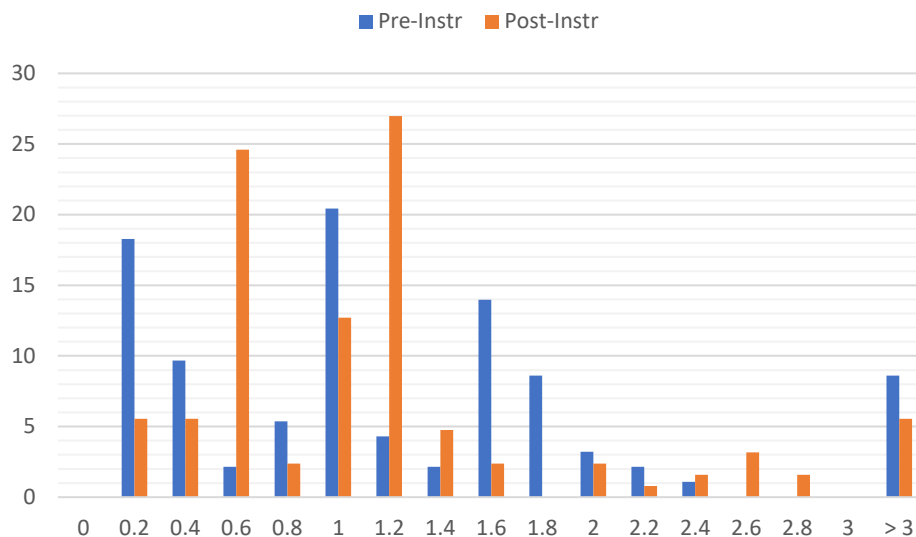
student ranges and the combined expert range), a smaller value represents less information loss, and so better agreement. In the pre-instruction survey, for example, the expert range was  $4.68 \pm 0.06$  inches. A common (but excellent) student estimate was  $4.7 \pm 0.1$  in, which has a resistor distance of 0.15. The (less expert-like) half-increment result discussed earlier ( $4.75 \pm 0.25$  in) gives a result of 0.9. This is still in the better half of student results; even though the uncertainty is large, the length estimate itself is close (within two expert sigma) of the expert range. The large spike of student values at K-L value of 1.6 are mostly the students who used the full increment instead:  $4.75 \pm 0.5$  in. Values of the K-L distance greater than 2 are students that chose very small uncertainties far from the expert value; for example,  $4.55 \pm 0.01$  in. The K-L resistor distance is largely insensitive to where the average is located if the uncertainty is extremely large: for example,  $4.75 \pm 0.5$  in and  $4.5 \pm 0.5$  in are assigned values of 1.58 and 1.64, respectively. This metric assigns results in a way that was largely intuitive: small values are reserved for close averages with similar uncertainties, and even an excellent average estimate doesn't make up for a very large uncertainty.

Overall, there were a few differences between the pre- and post-instruction distributions. First, there were fewer of the most expert-like estimates, whose K-L resistor distance was less than 0.2. This is in part due to the length estimates, not the uncertainty estimates. The expert range for the post-instruction survey was  $3.62 \pm 0.1$  in, but the most common student length estimates were 3.5 or 3.75 in which are more than a standard deviation away from the expert result. The post-instruction arrow was intentionally not lined up with the edge of the ruler, with around 0.1 inches of space on the low end of the ruler. This likely led to both the larger uncertainty estimates from the experts (since the lineup of the ruler is harder to gauge) and the frequency of the 3.75 length



estimate (which is a reasonable estimate of the far end of the arrow). In particular, the aggregate expert uncertainty of 0.1 in was somewhat larger than the pre-instruction survey's 0.06 in.

**Graph 4-3: Percentage histogram of comparing student and expert ranges (percentages shown), for both pre- and post-instruction surveys.**



These two effects push the K-L resistor distance in different directions: the distance tends smaller because the expert uncertainty is higher (and thus closer to the typically higher student uncertainties), but the distance also tends higher since fewer students estimated an accurate average. Using the average length and uncertainty for each class, one can compare the aggregate performance of the class to the expert distribution using the K-L resistor distance as well. All together, the post-instruction performance of the class is quite a bit better (0.5 vs 1.0), but this is mostly because the student errors in estimating the length averaged out. The aggregate student length of 3.68 in was therefore close to the expert value of 3.62 in, even though only seven students chose lengths between 3.7 in and 3.5 in. And though the average student uncertainty was nearly

the same (0.26 in pre- vs 0.28 in post-), the increased expert uncertainties led overall to a much smaller K-L distance. It makes more sense, then, to consider individual student results, because aggregating the results can elide individual errors in estimating the length.

Overall, the pre- and post-instruction groups behaved similarly in comparison with the experts via this metric. Although the post-instruction surveys more often overestimated the length of the object, the overall resistor distance was mitigated somewhat by the larger expert estimated uncertainty.

#### **4. TA dependence of the resistor distance.**

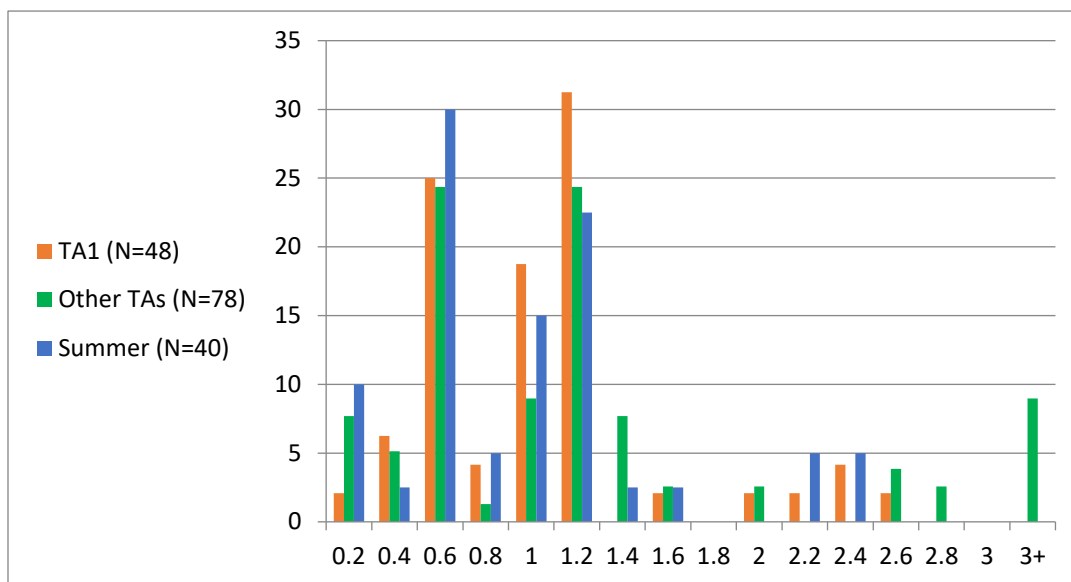
Although the entire 1BL lab class experienced the new uncertainty pedagogy at the same time, implementation details likely varied since the individual lab sections were taught by four different teaching assistants. While each TA was responsible for attending weekly lab training (see Chapter 2 for more details) where the techniques of uncertainty measurement and propagation of errors were emphasized, the teaching assistants were the ones communicating these changes to the students through both in-class guidance and lab report grading. Since only some of the TAs produced expert-like estimates of the length and uncertainty of the arrow on the surveys they took (see Table 4-1 for the TA's estimates) , it is certainly plausible that their students might have experienced a different quality of instruction on the same instructional materials and techniques, especially pertaining to uncertainty. Besides this, one of the teaching assistants (hereafter "TA 1") had, of their own accord, approached the LTACs interested in possibly making changes to the uncertainty part of the class, not knowing that this plot study was planned. In addition to this unusual motivation, TA 1 estimated the arrow in an expert-like way.

Indeed, students of TA 1 were more likely to be “sure estimators” than the students of other TAs after instruction. A full 40% of TA 1’s students were sure estimators, as opposed to only 22% for the other TAs (significant difference at  $p=0.015$ ). Pre-instruction, the aggregate percentage of sure estimators was 16% for either group.

A similar distinction can be found in examining the distribution of K-L resistor distance values in comparing TA 1 to the other three TAs, as seen in Graph 4-4. In comparison to the other TAs, TA 1 had more moderate values in between 0.6 and 1.2. 54% of TA 1’s responses fell into that range, as compared to only 35% for the other TAs ( $p=0.015$ ). This range was typified by uncertainty estimates such as  $3.75 \pm 0.05$  in ( $K-L = 0.93$ ) and  $3.7 \pm 0.5$  in ( $K-L = 1.1$ ). (The expert range was larger for the post-instruction survey, at  $3.62 \pm 0.1$ , which allows for even large uncertainty values like 0.5 to result in moderate K-L divergences around 1 if their associated average is similar to the expert result)

By contrast, the other TAs had more students on the high end of the range. 28% of their students had values greater than 1.2, compared to only 13% for TA 1 ( $p=0.02$ ). At the low end, these range estimates are typified by large uncertainties along with extreme averages, such as  $4 \pm 0.5$  in ( $K-L = 1.4$ ). The highest K-L values come from results with very small uncertainty estimates coupled with too-large size estimates compared to the expert values; for instance,  $3.8 \pm 0.01$  in ( $K-L=3.5$ )

**Graph 4-4: Comparing K-L resistor distance by lab instructor for post-instruction surveys. (Summer follow-up study taught by the author for comparison)**



In addition, TA 1's students were much less likely to estimate very large values of the arrow length (like 4 in), which made it basically impossible to construct a good overall estimate of length and uncertainty, whatever the choice of uncertainty value. (Only one of TA 1's students estimated the arrow length as 4, in contrast with 13 students for the other three TAs)

Taken together, TA 1 students generally did not perform at an expert level, but were less likely to choose extremely incorrect values for both length and uncertainty. I found similar results in my own summer class, working with the same arrow diagrams. Post-instruction only 15% of the students had K-L values greater than 1.6, and for the same reasons. These results point to the importance of better TA training surrounding appropriate estimation in the class: only the author and TAs that were most invested in the class saw this drop in extreme outliers for both length and uncertainty estimation.

The K-L resistor distance method has proven to be a useful tool in combining value and uncertainty estimates of students in comparison to the expert distribution. At least in these data

(where length values fluctuated by around 25% around the expert mean, but where uncertainty values could be different by orders of magnitude), it helped us to distinguish between different kinds of estimates in a quantitative way, especially in regimes where a raw overlap approach would have yielded nearly zero overlap.

### **Follow-up study**

In the pilot study, student uncertainties were seen to cluster strongly on specific values, as shown in the pre-instruction histogram in Graph 4-2; 87% of student values were 0.01, 0.1, 0.25 or 0.5 inches. These are the same numerical values that would arise from the following non-estimation procedures: the minimum measuring increment, in this case 0.5; half of the minimum measuring increment, in this case 0.25; one increment in the smallest decimal in the length measurement, in this case either 0.1 or 0.01 (depending if the length estimate had one or two decimal places).

The first two types of estimates are rules of thumb that are based only on the measuring device alone, and will be inappropriate when the measured object is difficult to measure (e.g., a round object measured with a straight ruler) or when the measuring device itself has systematics or other issues. In the case of the arrow measurement; the wide spacing between marks on the ruler was chosen in part to play up the absurdity of using only these rules to determine all uncertainties.

However, simply examining the value of the uncertainty that a student chose is not the same actually knowing their method of determining that value. Are the students who chose large

values of uncertainty like 0.25 inches merely poor estimators, or are they actually applying the half minimum measuring increment rule?

Certainly, the large values of uncertainty likely indicate that those students (or teaching assistants) were not examining the edges of the range that those uncertainties imply. Even if an expert's initial instinct (for measuring the arrow in Figure 4-2) was to use half the measuring increment, say, as an estimate of the uncertainty, a final answer of  $4.4 \pm 0.25$  in implies a range between 4.15 – 4.65 in; expert sense-making recognizes that the edges of these ranges are certainly not plausible values of the length of the arrow. In particular, the upper limit of 4.65 is certainly a larger number than the hash mark at 4.5 in; the arrow clearly does not extend to that mark. An expert check on the implied range end points therefore indicates that a smaller uncertainty is warranted (or, at least, a hard bound on the upper limit at 4.5 in)

Expertise in this task consists of using some sort of spread-based approach such as examining the plausible upper and lower limits of the length of the arrow, and then determining a plausible final value for the uncertainty. Thus, student expertise must not be simply evaluated on the number of their uncertainty; two students who both think the uncertainty is 0.25 in may be at different skill levels, if one is estimating (a too large number) based on the range, where the other is applying the improper rule of always using half of the minimum measuring increment. Class time spent on how to assess one's own estimates might be useful for the former student, but not for the latter who is correctly applying an inappropriate technique.

In a follow-up ruler measurement study performed in Fall 2012 with 1AL students, students were asked to describe how they determined their value of the uncertainty on the same ruler measurement questions as in the previous quarters. Student responses were binned according to the rubric in Figure 4-3.

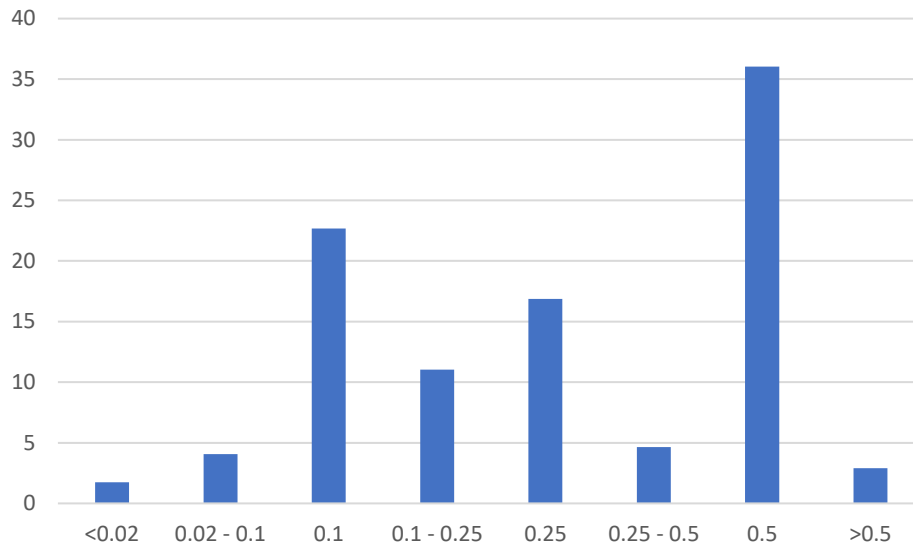
- a) Minimum ruler increment
- b) Equal to the last known significant figure from the estimate or ruler
- c) Reports length from estimated average to the nearest ruler increment (either larger or smaller)
- d) Measured the length multiple times
- e) Explicit estimation
- f) Established range, then calculated uncertainty
- n) I don't know / nothing
- o) Other

**Figure 4-3: Rubric for student uncertainty techniques.**

Using this rubric, categories d), e), and f) were tentatively considered spread-based, with c) possibly transitional.

Although this measurement question was asked both pre- and post-instruction, there were data collection issues in the pre-instruction survey that limited the amount of data collected. This lack of raw numbers (N=47 total) limited the usefulness of breaking down the data by each technique. These problems were addressed for the post instruction survey (N=172), so that survey is used for the following data. The expert assessment of the arrow used in the post-instruction survey was  $5.15 \pm 0.07$  in, with the arrow shape clearly resting in between the 5 in and 5.25 in increments.

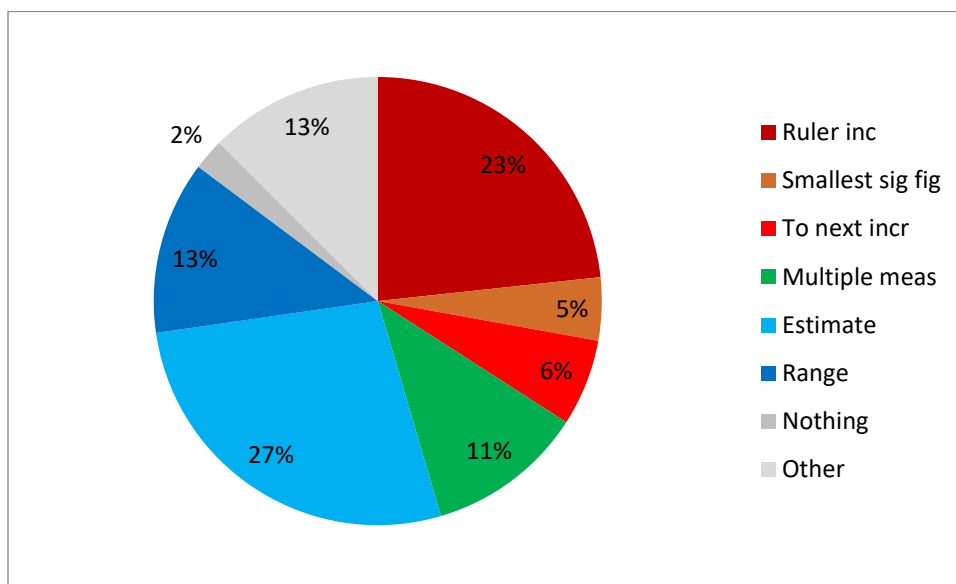
**Graph 4-5: Percentage histogram (N= 172) for Fall 2012 1AL post-instruction arrow uncertainties.**



Looking at the raw histogram of the data in Graph 4-5, the uncertainties determined by the students are similar to those in previous studies: the student answers are dominated by 0.1, 0.25, and 0.5 inch estimates. The number of “sure estimators” in this data is 20%; in the previous study, these were assumed to be evidence of estimation. Student self-reported methods for determining uncertainties are shown in Graph 4-6. (Their techniques have been classified according to the rubric in Figure 4-3).



**Graph 4-6: Self-reported methods for determining uncertainty (N=172).**

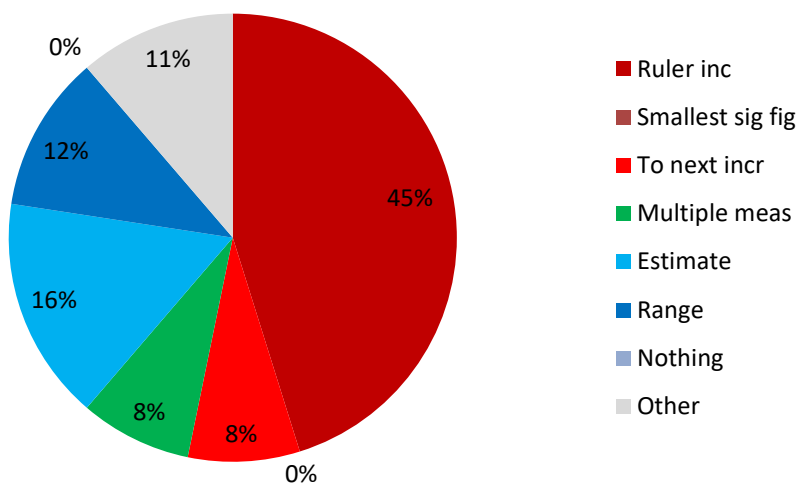


In that chart and the following method charts, concrete techniques pertaining to the ruler (using the ruler increment or half of the increment directly; the smallest significant figure from the length estimate; and measuring a distance from the end of the arrow to the next ruler increment, respectively) have been colored in different shades of red. The two blue shades indicate explicit estimation or establishing a range of possible values and determining an uncertainty from that range. The green indicates a student making multiple measurements (usually drawing multiple lines between the arrow and the ruler) and determining the uncertainty from that. We also consider that technique spread-adjacent, so the blues and green indicate those approaches. Finally, grey colors indicate unclassifiable or no explanation (but still a numerical uncertainty estimate).

In aggregate, spread-based techniques account for around half of the total, including 27% for estimation alone. Using the smallest ruler increment is the next most commonly used technique, at 23%. The initial expectation was that ruler increment methods would be the dominant

approaches for students that chose 0.25 or 0.5 in uncertainties, while estimation would be more common for the intermediate bins, like values between 0.1 in and 0.25 in.

**Graph 4-7: Self-reported methods for determining uncertainty for 0.5 in values (N=62).**



Indeed, the ruler increment approach was the most common method by far for the 0.5 in responses, as seen in Graph 4.7. Nearly half of the students who had that uncertainty simply marked down the smallest increment of the ruler that they saw. If the arrow had been placed next to a millimeter ruler, they likely would have chosen a 1 mm uncertainty, which would represent a much more reasonable value. Since this survey item was chosen to play up the contrast between this approach and a reasonable length estimate, this exaggerates the problem with the minimum increment method compared to real life applications.

More surprising was the significant minority of estimators or range measurers that still recorded such a large value of uncertainty; this value of 0.5 inches is five times as large as the

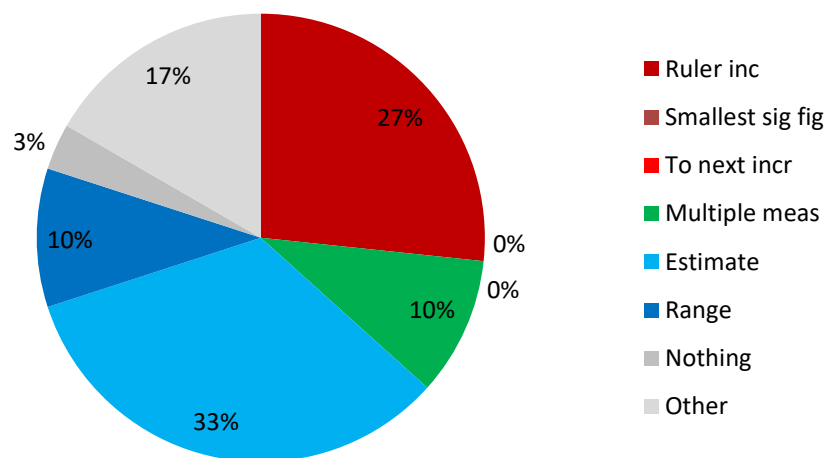
expert value of the uncertainty for this arrow diagram. This likely represents the lack of a “sanity check” on the estimate or, possibly, thinking that the uncertainty meant the full size of the range rather than half of the range.

As expected, most of the rest of the ruler increment measurers chose a 0.25 in uncertainty, since the other common increment approach is to take the minimum ruler increment and divide by 2. Here, though, the estimation approach is an even larger fraction of the total, yielding a third of all responses (as seen in Graph 4-8). While a 0.25 inch estimate is still large, it is only around twice as large as the consensus expert value. Still, this result is too large to be consistent with the understanding of the uncertainty as one side of the full range. For example, a common response was  $5 \pm 0.25$  inches. While the 5.25 in side of that range is perhaps a reasonable upper limit, the other side of 4.75 in is not, considering the shape clearly extends past the 5 inch ruler marker. This is a typical difficulty in dial measurement problems, where there is usually either a clear upper or lower limit. To readjust the size of an uncertainty to reflect that limit (but being careful to make it make sense on the other side of the range as well) is certainly an expert-level estimation skill, especially on the fly. The non-expert graduate student responses were also of this type.

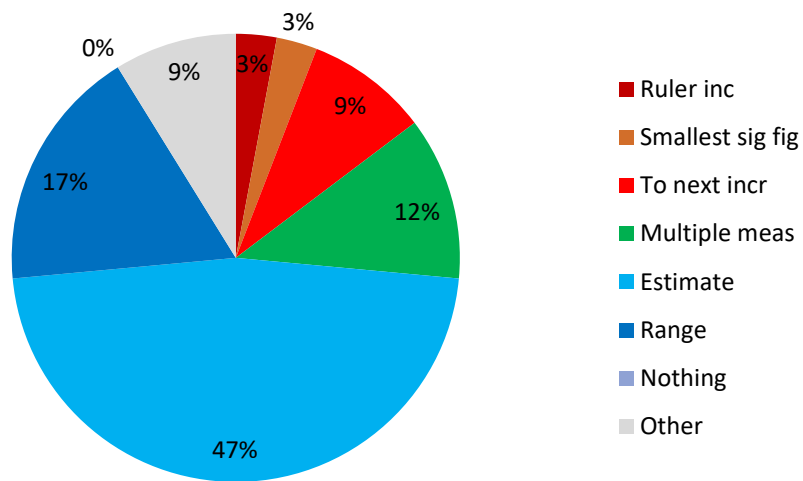
Intermediate “sure estimator” ranges (such as values between 0.1 and 0.25) were even more heavily dominated by estimation approaches, as seen in Graph 4-9. For these in-between values, 70% of students used a spread-based or estimation approach. When the “sure estimator” heuristic was originally formulated for the pilot study, it was suggested that students may have been measuring from their length estimate to the nearest ruler increment (scored in the charts as the bright red color). For example, a student that estimates a length value of 5.2 inches might report an uncertainty as 0.2 inches (or 0.3 nches) as that is the length to the next ruler increment. Unless the length estimate lined up exactly on an increment (or exactly halfway between two increments),

these students would always produce one of these intermediate uncertainty estimates and would have been sorted as “sure estimators” in the pilot study, despite using a ruler increment-based approach. But here we can see that these students make up only 9% of the methods for these in-between uncertainty values, so summarizing this whole category as composed up of estimators or other range reasoners turned out to be basically correct.

**Graph 4-8: Self-reported methods for determining uncertainty for 0.25 in values (N=29)**

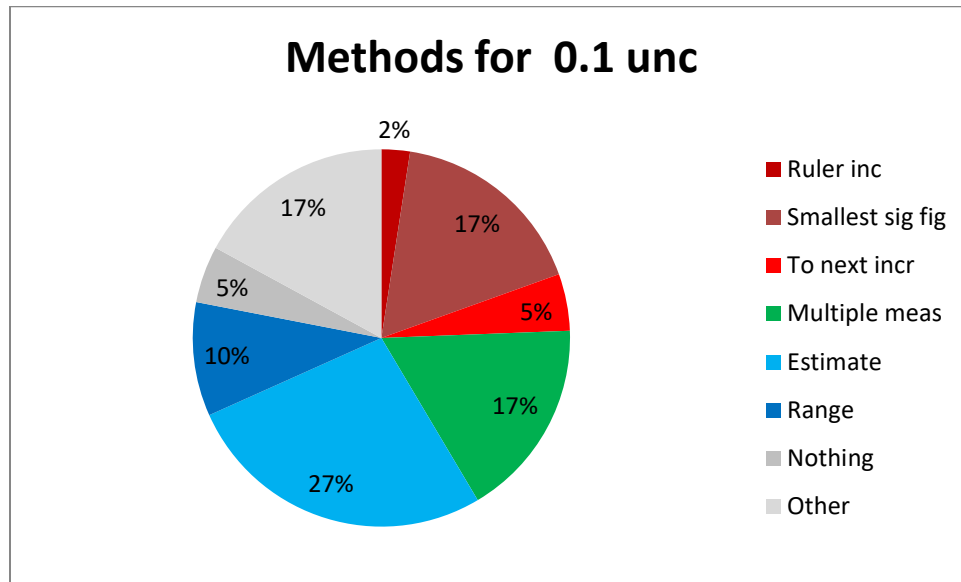


**Graph 4-9: Self-reported methods for determining uncertainty for values between 0.02-0.1 in; 0.1-0.25 in, and 0.25-0.5 in, aggregated (N=34)**



Students choosing a 0.1 inch uncertainty displayed the most diversity in their methods, as seen in Graph 4-10. Most of these students had a one decimal place estimate of the length, and so the smallest significant figure method would lead them to report a 0.1 inch uncertainty; 17% of the 0.1 inch responses used this method (Interestingly none of the 24 students with a two decimal place length estimate chose a 0.1 inch uncertainty, despite it being the second most common uncertainty value overall). Still, the plurality is still direct estimation; spread-based approaches still make up the majority.

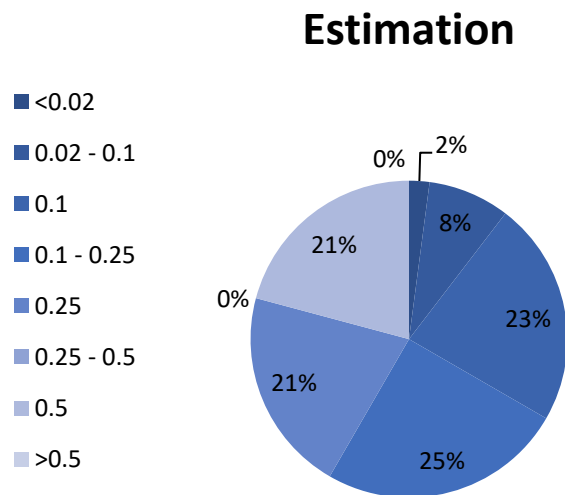
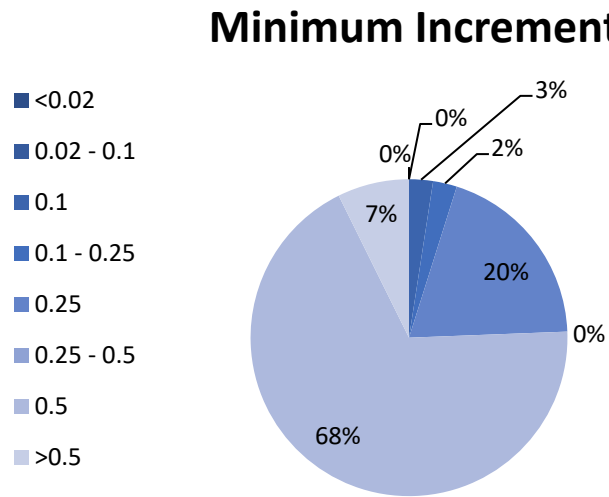
**Graph 4-10: Self-reported methods for determining uncertainty for 0.1 in values (N=39).**



Another way to examine the data is to look at the methods themselves, and see which uncertainty values most commonly resulted from each measurement approach. This shows whether certain approaches are more likely to give reasonable values of the uncertainty when used by the students.

As expected, use of the smallest ruler increment technique leads overwhelmingly to uncertainties of 0.5 and 0.25 inches, which make up almost 90% of those responses. (A minority of students interpreted the minimum ruler increment as being 1 in, which accounts for the results greater than 0.5 inches) This method yielded an average uncertainty of 0.47 inches, five times greater than that of the expert TAs. This divergence is natural, since the ruler (with few increments) was designed to exaggerate the differences between this approach and measurement expertise.

**Graph 4-11: Uncertainty values for the minimum increment method (N=41) and the estimation method (N=48).**



By contrast, students who explicitly estimated the measurement uncertainty had a much more diverse selection of uncertainty values, reflecting a difference in the quality of estimates. Unlike a ruler increment method, which can only yield a few different possible results if the machinery is used correctly, estimates can a variety of values, some of which are too large and

others too small. Given the way the arrow measuring problem was constructed, non-expert estimators would likely estimate numbers that were too large rather than too small. In aggregate, the estimators average 0.25 inches for their uncertainties. These estimates are dominated by four different values: 0.5 inches; 0.25 inches; between 0.1 and 0.25 inches; and 0.1 inches. No estimators chose values in between 0.25 and 0.5 inches or greater than 0.5 inches. This might indicate a tendency for students to estimate small values for a given decimal place (e.g. 0.2 vs 0.6), because larger value estimates “should” be rounded up to the next decimal place or measuring increment. This would tend to round those large intermediate estimates to either 0.5 or 1.0 inches. The 8% of students that estimated uncertainties between 0.02 and 0.1 also chose values equal to or less than 0.05 inches, providing more evidence for that hypothesis.

Actual statistical calculations of the standard deviation or of 95% confidence limits, of course, can produce any value for the uncertainty. Ideally, a robust estimation technique could also yield these sorts of values, to better approximate the real ranges of these quantities. Two of the five expert TA estimations (see Table 4-2, in the previous section) were 0.03 and 0.0625, which are in the range that no student estimator chose. Interestingly, the other two spread-adjacent approaches (multiple measurements and range; colored green and dark blue respectively in Graphs 4-6 through 10 above) did include values between 0.25 and 0.5 inches; these students represented between 10-15% of those methods.

While the relatively small number of respondents for each of these methods (N=20 and N=22) prevent confidence in the detailed category statistics, they both display a diversity of values similar to the estimation approach. These methods yielded average uncertainty values of 0.24 in (multiple measurements) and 0.33 in (range). Both methods had a large minority of 0.5 in values that raised the average, contrasted with more expert-like smaller estimates. The measurement



technique that yielded the most expert-like uncertainty value was the smallest digit of the length approach; this, however, is just a coincidence, as the smallest digit approach would give 0.1 inches for one decimal place average estimates. This happened to be near the expert uncertainty value for this picture, but was not for the pre-instruction arrow<sup>10</sup>.

All in all, the method of determining uncertainty had a strong impact on the actual uncertainty value, especially since several of the approaches yield concrete, repeatable results when applied to a variety of different situations. An expert approach should not be so context-free, however; distributions have standard deviations or spreads that can take any value, not just half of the measuring increment! Even if students estimate or use a range, however, their estimations can lack the internal feedback process of assessing and reassessing that an expert performs. In conversations in lab or in office hours where students needed to estimate something, I often found that they would start by saying they didn't know; then, when coaxed, choose a value that didn't fit the context of the situation. In most cases, explicit guidance by the TA was required (e.g., audibly examining the upper and lower limits for a given students' uncertainty estimate as a sanity check) before students would choose estimates that were more defensible. This mirrors the internal process that experts undergo as they estimate; making this explicit allows students to realize that an expert estimate isn't simply a genius perfect guess, but a balancing of multiple factors, often thought through quietly and quickly. In this sense, expert estimation is a range or spread-based approach, as the individual edges of the range implied by the estimate are judged reasonable before the final estimate is chosen.

While this sort of scaffolded estimation exercise was only informally used (and then, usually only in the early 1AL measurement lab), I think it is essential for helping students transition

---

<sup>10</sup> It is a challenge to devise an arrow measurement problem whose uncertainty avoids all of the possible ruler-based techniques!

away from the concrete increment-based techniques to estimation or range-based approaches; and from being inaccurate estimators to being accurate estimators. Although the primary focus of our changes in 1-Series lab curriculum was on consistently recording measurement uncertainties, propagating those uncertainties to unmeasured quantities, and on comparing results based on those uncertainties, a larger focus on determining uncertainties via estimation is a natural expansion that it makes sense to pursue in the future. See Chapter 5 for more on future directions of uncertainty projects.

### **Why measure multiple times in lab?**

One positive thing that cookbook labs do tend to emphasize is multiple measurements of the same quantity. Several positive outcomes could stem from such a practice – increased procedural practice; more data leading to easier recognition of null results; etc – but the most important in an uncertainty focused class is that it allows the determination of a measurement uncertainty at all. With only a single measurement of each quantity, the precision of the measurements cannot be determined except by convention (e.g., significant figures).

Although the 1-Series lab classes did not explicitly connect the lab instructions (measuring multiple times) to the calculation of uncertainty, students may have picked up this understanding as they proceeded in the class. After all, during each lab they dutifully calculated a measurement uncertainty from multiple measurements of a single quantity, and then repeated this process for virtually all measurables. And in those rare cases where multiple measurements were impractical, they estimated an uncertainty for the quantity.

During three lab semesters (both pre- and post- instruction), students were asked to describe why they would want to measure something in lab more than once. The focus of the analysis was tracking how often student answers contained a reference to establishing the uncertainty or spread of the data. The fractions shown in Table 4-3 are the fraction of responding students who had at least one spread or uncertainty-based answer in their response. Common non-spread reasons for measuring multiple times included: checking procedural mistakes; calculating a better average; and more generic spread-adjacent responses like “for accuracy” or “for precision”. (Those specific two word responses were frequent, but were not counted as uncertainty-based because of their vagueness) In some cases, the spread-based responses were quite short (“to calculate uncertainty” for example), but still represented use of the class’s language around uncertainty and error. Often these responses would be coupled to finding a better best estimate: “to get a range of possible answers. Doing it multiple tries helps give a more accurate number since we can take an average.”

In all cases, the number of students choosing these uncertainty-based responses increased after instruction. For all semesters except 1BL Winter 2013, the students were in their first semester of the new pedagogy. The Summer 2012 class did not show a significant difference, but that may be because of the much smaller number of summer students compared to the other three semesters. The Spring 2012 data’s extremely low p-value is suspect, since there were only 5 students in the pre-instruction sample that chose a spread-based response. Still, that number ballooning to 25 students post-instruction likely represents a real difference.

**Table 4-3: Fraction of students responding that they made multiple measurements in lab to determine uncertainty or spread of data. Significant p-values are bolded.**

Class	Pre-instr.	Post-instr.	p-val	Cohen's h
1BL (Spr 2012)	0.02 (N=138)	0.17 (N=141)	<b>&lt;0.001</b>	0.49
1BL (Sum 2012)	0.23 (N=43)	0.31 (N=45)	0.20	0.18
1AL (Fall 2012)	0.17 (N=46)	0.32 (N=169)	<b>0.03</b>	0.34
1BL (Win 2013)	0.21 (N=86)	0.35 (N=97)	<b>0.02</b>	0.32

The students in the Winter surveys were further divided into two groups: those that had previously taken the modified version of 1AL in the Fall, and those that had not. Fall 2012 was the first 1AL class that had used the new lab pedagogy, so students could not have taken a different modified 1AL lab class. The expectation was that students that had taken two semesters of modified lab classes might be more easily able to recognize that multiple measurements were largely taken to determine a measurement uncertainty. Indeed, students that had taken two semesters were more likely (0.38 vs 0.22) to use a spread-based approach. This is a suggestive result rather than a statistically significant one, however; only 18 of the 97 respondents hadn't taken the modified 1AL course, limiting the power of the statistics. Similar results were found in the pre-instruction survey for Winter. Of the twelve students who answered the survey that did not take that fall class, none of them used a spread-based response.

At minimum, this change is reflective of students beginning to use the language of the class (“uncertainty” and “spread”) to describe their results. This is ultimately a success of the lab frame: the focus on how multiple measurements were used (i.e., to help establish a spread for

comparison's sake) has seemed to transfer into the students' sense-making for this task, as opposed to remaining an arbitrary task that they perform because "that is what you do in a lab."

More broadly, the entire project was centered on helping students learn how to measure, calculate, and draw appropriate conclusions from their data. Of course, these goals are hardly possible on a quantitative level without a focus on measurement uncertainties. As described in more detail in Chapter 2, each conclusion prompt (and many of the in-class checkpoint questions) was centered on determining and propagating uncertainties and using those uncertainties to make comparisons between different data sets.

While it was obvious to the students that the class focused on uncertainty, the reason of this focus was less explicit. For that Winter 1BL class, students were also asked why they thought we asked them to measure uncertainties in the lab.

- a) Procedural mistakes / errors (in-lab measurement mistakes)
- b) Descriptions of the world: "everything is uncertain", data might be wrong
- c) Accuracy or precision (either "for accuracy" or "to improve precision")
- d) Range or spread of data
- e) Range or spread of data, but also saying that that range indicates confidence in results
- f) Uncertainty needed to compare results with predictions or other data
- g) Uncertainty needed to "account for error"
- h) Allow error to be calculated
- i) Part of becoming a scientists/future application
- j) Other

**Figure 4-4: Rubric for why the class focused on uncertainty.**

A rubric was crafted based on the student responses to that survey item. Because of the open-ended nature of the survey item, students answered the question using a variety of different interpretations. The rubric categories in Figure 4-4 themselves can be further sorted to represent

the different types of approach to the question that students used. Quotations in the following sections are from the student surveys.

#### Category 1: Uncertainty as an unfortunate, inevitable mistake (a&b).

The most basic understanding of uncertainty is as error(s): literal mistakes in measurement that prevent the correct answer from being determined. These can be viewed as mistakes of the measurer that are in principle avoidable but in practice are not because humans are imperfect. This is the origin of lab conclusion bogeyman “human error”, meaning an unquantified but unavoidable mistake that explains whatever incorrect result that was found. Even more broadly, these errors can extend to the measuring devices or mechanisms (“so we can be wary of all possible existing errors and flaws of an experiment so we can understand and correct them”).

#### Category 2: Uncertainty as a data feature (c&d&e&g&h).

These classifications focus on uncertainty as a feature of the data or measurement. This category represents the continuum of understanding mentioned in Allie, et al. [6]: from nonspecific accuracy or precision to a numerical quantity related to the range of measurements. For c) and g), uncertainty represents the data in some way, even if the procedure to determine it is not specified. Every data set *has* some precision or accuracy (c); or its uncertainty can be accounted for in some way (g). The use or value of this is left unspecified for these codings, however.

The h) classification connects that accounting to some calculational procedure: uncertainty as something that is the end product of some math operating on the raw dat. While an appropriate

lab frame perception (indeed, for virtually every lab students do calculations to determine measurement uncertainties), it omits what the procedure really calculates (e.g., “to see what is the value of the error”).

This connects conceptually back to category 1 and the ideas of uncertainty as mistake: the mistakes (whether human or environmental) broaden the idea of a single ideal measurement to a range which encompasses those errors (d). One student expressed the uncertainty as a collection of mistakes around a true value: “[to] take into account slight variations from actual measurement.” Data with many mistakes or unavoidable errors yield a larger range, which summarizes the quality of the experiment (e). These begin to edge into the ways that uncertainty can be useful, as it “helps us determine whether our conclusion is reasonable.”

### Category 3: Reason to study uncertainty (f&i)

The above progression describes how uncertainty represents the data, but not why how it is helpful or useful. Ultimately, the pinnacle of lab-frame understanding of uncertainty is realizing that uncertainty is a necessity to compare data (f). To determine if two sets of data are consistent, they must “fall into range of each other.” Without an uncertainty, an experiment is not reproducible or reconcilable with other experiments or with theory. This is really why measurement uncertainty is taught: it is a requirement for doing useful science (i). As one student put it, uncertainties are measured in lab “so that we can draw correct conclusions based on our data and so that we learn to be scientists.”

Even at the end of one (or two) semesters of instruction, few students responded in ways relating to these more abstract purpose of the uncertainty instruction (7% f responses, 6% i responses). This makes sense, as most of the course in practice was oriented towards teaching the mathematical procedures for measuring, calculating, and using uncertainties rather than the why. See Chapter 5 for some examples of possible course alterations to address broader uses of uncertainty.

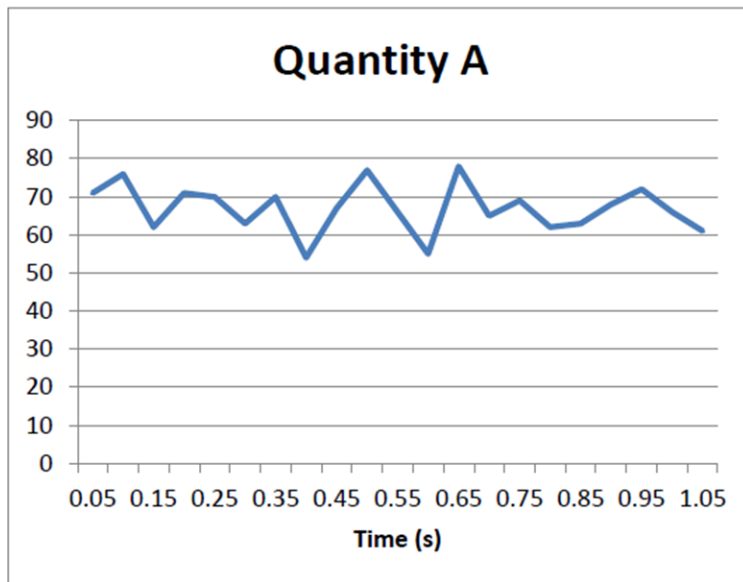
There were not any statistically significant changes of number in any response category between pre- and post-instruction. The number of survey non-responses did shift from 32% to 15% ( $p=0.002$ ), perhaps indicating increased confidence in answering this question after being reminded of lab tasks during the semester.

### **Determining uncertainty from a graph**

Computer data acquisition systems (such as LoggerPro) are frequently used in 1-Series labs, as a way of automatically collecting larger amounts of data than would be practical by hand, as well as making use of digital sensors. Typically, the computer display showed a graph similar to the picture in the survey item in Figure 4-5, showing the random fluctuations of a measured quantity. To determine the value of the quantity, students sometimes used the built-in mean and standard deviation functions, and other times determined the uncertainty from the graph directly, by determining the upper and lower limits, and working out the measurement uncertainty from there. Although the ability to calculate the statistics at the press of a button is convenient, that method functions as a black box, especially since there is no description of what the standard



deviation actually means. This is similar to instructing students to draw a best-fit line by hand rather than allowing Excel to do it for them: while the actual data product is less accurate, it helps to illuminate what the statistical product actually means and what relation it has to the data.



You use a Logger Pro-like data acquisition system to make the following measurements of Quantity A, as shown in the picture above. Using this data, answer questions 2, 3, and 4:

2. How large is quantity A?
3. What is the uncertainty in this measurement?
4. How would you record this measurement if it were required for a lab conclusion?

**Figure 4-5: Determining uncertainty from a graph survey items.**

In the survey, students were only able to use this latter means of determining the mean and uncertainty of the data shown, since there were no automated buttons to calculate the line's

statistics. Even though these tasks are superficially similar to the ruler measurement tasks discussed in the pilot study, they differ in a few significant ways. First, this task is essentially a lab task, unlike measuring a length with a sparsely-lined ruler. This survey object was provided to students during the first weeks of 1BL lab class, so they had a full course of experience in interpreting this sort of digital acquisition readout. (In 1AL, several labs in a row use digital sensors to measure distances, accelerations, forces, etc.)

Second, computer data acquisition may lend itself more easily to spread interpretations in the first place. Unlike direct human measurements, where each separate measurement might take tens of seconds, LoggerPro's distance sensor (for instance) typically operates at around 20 Hz, leading to hundreds of data points in a matter of seconds, all measuring the same system. Here, differing measurements cannot be as easily swept under the rug as "human error"; instead, they are an inevitable and quantifiable result from the process of measuring. Consequently, students may be more easily primed to consider measuring range for these systems rather than for a handful of direct human measurements. Third, the task gives more potential insight to the lab frame experienced by the students. Consequently, we included the last question of the item ("how would you record this measurement if it were required for a lab conclusion?") to put the students as cleanly as possible in their lab writeup frame. This allowed us to see whether student self-reporting of the measurement would include the uncertainty as well as the average.

Finally, even the initial measurement of the quantity may make use of some range techniques. Unlike a ruler measurement exercise, where the mean is easily determined by lining up the object to the ruler and reading off the value, determining the mean by looking at a graph implies looking at the whole data set at once.

Besides determining the average and uncertainty, students were asked how they would report that information in a lab report. This was a way of testing a basic lab action: the recording and reporting of digital data. 64% of students responding to the question reported their results together as an average  $\pm$  uncertainty. Of those students, 74% expressed a numerical range that was plausible, given the data; the others either doubled the uncertainty range (or literally wrote “average  $\pm$  uncertainty”).

Of course, this question represents the maximum possible priming with respect to spread that students are likely to encounter: because the previous survey items ask specifically for average and uncertainty of the data, the high rate of spread-based reporting is no surprise. (Indeed, in this it is similar to the priming in the conclusion prompts and checkpoint questions, which would always specifically refer to the uncertainty for some measured quantities.) Other respondents reported that they would display the data in a table (the most common form of data display in the lab manual) or described the steps they would take with the computer data acquisition software (“I would find the standard deviation by highlighting all the data, then use that to find the uncertainty”).

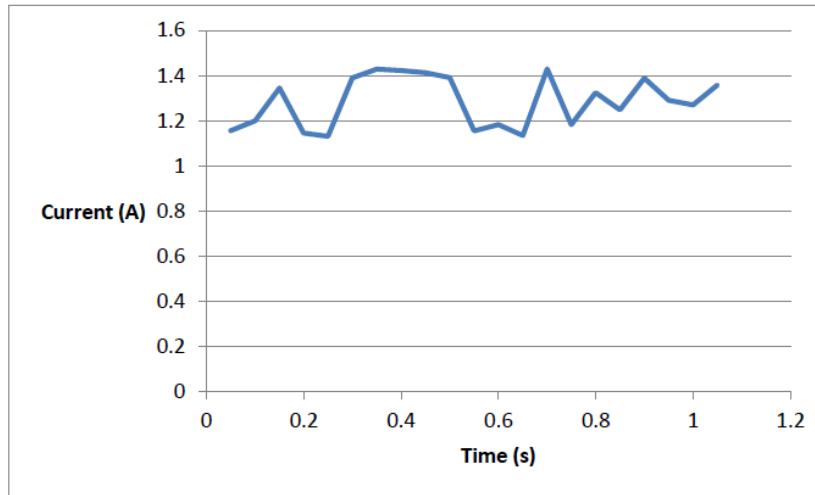
Interestingly, a significant minority of students used a spread to represent even the average value of the graphed quantity: 22% expressed the size of the quantity as a range (e.g.: 55-79), with two students expressing it as an average  $\pm$  uncertainty. These students may actually be viewing the data in a spread-based way, rather than just responding to the uncertainty priming especially since this question was asked prior to the reporting data item. While this attention to the maximum and minimum values of a range of data is taught in the class (especially with respect to propagation of errors), the graph itself neglects the average (which cannot be ascertained immediately by examining the screen) while emphasizing the extrema, which are easier to line up on the graph.

This computer acquired data has the advantage of visually showing a spread, rather than relying on the students to infer multiple data points as representing a spread; this displays them as more of a part of the complete data rather than as “successful” and “failed” attempts to measure the same quantity. Further study is required to determine if this kind of display of multiple measurements would be more conducive to teaching the idea of spread than a list of data points.

Of the students that used a range to answer that question, only 4/19 skipped the subsequent question about determining uncertainty. For the remaining students, there are two uncertainty estimates: one explicitly stated as answer to the uncertainty question, and another implied by their range for the overall size question, by dividing the extent of their range by 2. Interestingly, only 5 of those 15 students reported the same uncertainty that their range implied. In some cases, these were simple errors, like quoting the full data range as the measurement uncertainty (e.g. a range of 60-80 with an uncertainty of 20). Only one student seemed to use the “measuring increment” as the source of their uncertainty: they estimated a range from 54.5 to 79 and wrote down an uncertainty estimate of  $\pm 0.5$ , the smallest increment of their estimate.

### **Graph uncertainty including propagation**

The post-instruction variant of this question was broadly similar, but also included a propagation element (see Figure 4-6). Rather than a completely abstract data graph, the data was framed as a measurement of changing current in a resistor, a common topic of 1BL lab classes. Though the students were shown simulated current data, they were asked how they would report the *voltage* across the resistor, having been also provided with a numerical value of its resistance.



You use Logger Pro to measure the current through a resistor. The resistor is assumed to be exactly 20 Ohms. Use the picture above to answer questions 3, 4, 5 and 6:

3. What is the current through the resistor?

4. What is the uncertainty of that current?

5. If your lab conclusion asked you to report the voltage drop across the resistor according to your data, what would you write?

**Figure 4-6: Determining and propagating uncertainty from a graph.**

Most students were able to estimate a reasonable number for the average and uncertainty of the current. 89% of the responses gave a numerical value of the current between 1.2-1.4 A, with the rest either choosing a current corresponding to the upper limit of the graph or reporting the resistance or voltage value (20  $\Omega$  or 26 V, respectively) instead. Only one student reported the answer to this question as a range; this was only 1% of the total responses, in comparison to the

22% that did so in the pre-instruction version of the same question. One possible reason for the discrepancy is the increased concreteness of the phrasing in the circuit version of the question: by emphasizing “current through the resistor”, that may have primed students to give a single numerical answer to correspond with that physical reality more so than “how large is abstract quantity A”.

The approximate range of the data was from 1.1-1.4 A. Consequently, uncertainty values between 0.1 and 0.2 A were considered reasonable for the data set, with larger or smaller values unreasonable. Even a value of 0.25 A more closely approximates the spread of the full range than the one-tailed uncertainty; 0.2 A was included as it represents a rounding of the expected 0.15 A range. 78% of students estimated an uncertainty within these bound. A further 6% chose values that approximated the full data range rather than the one-tailed uncertainty. This makes sense, since after (at least) one quarter of lab work, most students should be comfortable with the class’s definition and usage of the term uncertainty. Most of the remaining answers were very small, with the most common being 0.01 A. It is not clear how students determined these small estimates.

These results compare favorably to the arrow estimation problem, where more than half of survey students in each survey chose very large uncertainties (more than twice the expert estimate). A significant minority (23%) selected an answer in between the 0.1 and 0.2 A values (not including the endpoints), which is a strong signifier of range-based reasoning as opposed to a simple visual estimate of 0.1 or 0.2 A. This mirrors the calculation procedure taught in class: determine the upper and lower limits of the range, take the difference, and then divide by two to get the uncertainty. In this case, the upper limit is just over 1.4 A, and the lower limit is around 1.1 A. This makes the full range 0.3 A, leaving the uncertainty as 0.15 A. Although 78% of students gave a reasonable answer, that is likely an underestimate of the fraction of students that can functionally operate the

LoggerPro system to determine the uncertainty of an actual current sensor; here, there are additional barriers in estimating the upper and lower limits of the range that would be easily addressed with the display of the actual program, since those limits could be read directly instead of requiring some estimation.

For the final question, respondents needed to use their observation of the current's value and uncertainty, and calculate the corresponding voltage and its uncertainty, using Ohm's law and the technique for propagating uncertainty described in Chapter 2. In short, students should use Ohm's law with the extreme values of current to calculate the largest and smallest possible values of voltage; and then from these extrema, they could calculate the range of possible voltage values and then determine the uncertainty from that range. Although no equation was provided in the survey prompt, Ohm's law is the most memorable equation in 1BL. This survey was also given after the students had had three consecutive labs that made use of Ohm's Law, including propagating uncertainties with it. Indeed, of the 65 students that provided numeric answers to the question, only five (8%) calculated a voltage incorrectly (given their graphical estimate of the current) or wrote down a current instead of a voltage for the propagation problem.

**Table 4-4: Student reporting of voltage.**

Single voltage only	Average voltage $\pm$ Uncertainty	Range	Non-numeric	Total
27	27	6	28	88

The question concerning the voltage specifically asked what the students would write in their lab conclusion to report the voltage. Here, the key distinction was whether students would simply include their calculation of the voltage (their best estimate), or whether they would also include a propagated uncertainty. Because this survey item doesn't specifically ask for the uncertainty in the voltage, it records the students' own lab frame: what does it mean to report the voltage? Is it enough to write down the best estimate, or must the uncertainty also be included?

Of those that provided numeric answers, 45% reported a single voltage value, and another 45% included their average voltage  $\pm$  a numeric uncertainty. The remaining 10% expressed their value using the upper and lower limits of their voltage range. This is a strong result: more than half of these students calculated the spread of a derived quantity using their estimation of the measured quantity and their memory of the appropriate equation and propagation technique, and with only partial priming. Moreover, some of the students that did not calculate a number still expressed the voltage as the words "average  $\pm$  uncertainty", indicating that they knew a proper reporting of the voltage in the lab frame should include an uncertainty value.

It is important for students to propagate the uncertainty correctly, of course. The students that reported a range (either via the upper and lower limits or by listing an average  $\pm$  uncertainty) generally did so correctly given their assumptions about the uncertainty. Of these, 77% (23 of the 30) of students correctly calculated the range, with the remaining seven making various mistakes, mostly errors in calculation. The most common mistake was calculating one edge of the range correctly, but making an error in calculating the other edge of the range. Encouragingly, only one of the students did not attempt to propagate the uncertainty at all, instead using their estimate of the current uncertainty as their voltage uncertainty.



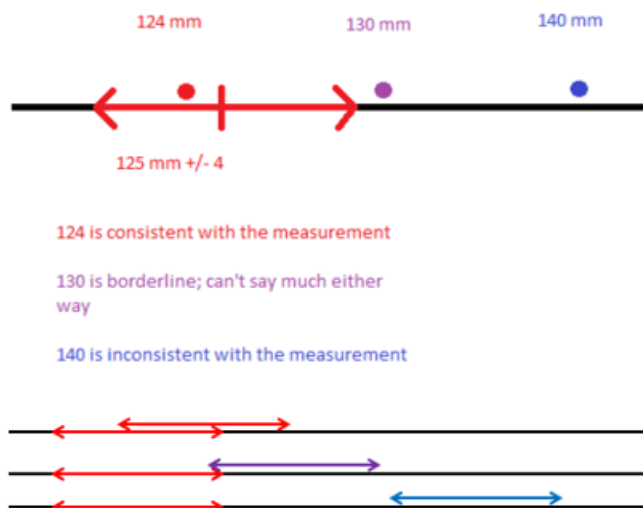
Even with the most conservative possible estimate based on the surveys collected (that only those 23 students could calculate the propagation of uncertainty, and each of the 124 other surveyed students could not), 17% of surveyed students propagated the uncertainties correctly without a reference. This certainly represents an underestimate of student competency in this skill in the class overall, since the students were lacking some of the resources (time, the relevant equations, data, access to the uncertainty document) that they would typically use to calculate the propagated uncertainty in lab conclusions. While we don't have any comprehensive data on how often students were able to complete the propagation procedure successfully on the actual lab conclusions, the students were broadly able to complete the procedure.

The ultimate goal was not for students to merely learn the procedure of calculating a range, but to use it to check the consistency of their data with respect to other data or a reference source. In a follow-up question, students were asked if their value for the voltage was consistent with 22 V, a value that had been determined by a student from another labgroup.

Responses were coded via a rubric (see Appendix V) to determine if the students used range or spread-based reasoning to determine if that voltage was consistent with their data. Just over half of general respondents (52%, 44/84) used spread-based reasoning, typically by determining if their spread included the 22 V value. Indeed, of the 30 identified above who reported the voltage as a range, 29 used spread-based reasoning arguments to determine if the value agreed with their range, a remarkable 97%. Therefore, we can be confident that the surveyed students that propagated uncertainties successfully were also comfortable using spread-based arguments to determine agreement, at least in the case of comparing their spread to a single reference value. As expected, the ability to propagate uncertainties (necessary to determine a spread in voltage in this case) is highly correlated with the use of range agreement methods in determining whether that

range is consistent with 22 V. After all, one requires an actual range of values to do spread-based reasoning in the first place. The correlation between the two is  $r=0.55$  (95% confidence limits from 0.46-0.62). Still, this means that students that calculated a range actively used it to determine if the 22 V was in agreement with that range. (There is an argument that student use of these techniques isn't precisely the use of spread-based reasoning, since the comparison is between a range and a single value, rather than two ranges. See the next section in this chapter on range agreement for further discussion, including how well students did compare two actual ranges to determine agreement.)

Even though virtually all of those students used spread-based approaches, they were split on whether the 22 V value agreed with their ranges; 16/29 agreed, 12/29 disagreed, with one not sure. 22 V was intentionally chosen as a borderline value: it is included in the range if one chose a larger estimate of the uncertainty of the current (0.2 A, leading to a lower voltage limit of 22 V), but excluded if a smaller uncertainty was chosen (0.15 A, leading to a lower voltage limit of 23 V). This emulates the kind of estimate-based variability found in the lab write-ups. Only two students total (one who propagated the uncertainty and the other who did not) said that they were not sure if 22 V agreed with their range. Most of the students, then, viewed the agreement as a clear-cut determination: either 22 V was contained within the range (and so it was consistent with it) or it was not (and so it was inconsistent). Even though in course materials these borderline values were explicitly mentioned, most students did not recognize them for what they were (See Chapter 2 for a fuller discussion). Proper interpretation of these borderline results also remained a difficulty in comparison of two ranges, as seen in the next section.



**Figure 4-7: Excerpt from uncertainty document discussing agreement of ranges.**

Future iterations of this coursework should further emphasize uncertain statements of agreement in borderline cases, as students seem to take a cut-and-dried view of overlap/agreement that does not map well to more advanced conceptions of measurement uncertainty (such as different number of sigma providing different levels of confidence). Perhaps ranking exercises (“which of these quantities are you most sure agree?”) can emphasize this specific area. Artificial set-piece problems might be necessary to approach this problem, since relying on student-taken data won’t always naturally lead to these sorts of edge cases. See Chapter 5 for further discussion of future topics.

### Comparing two ranges

One of the primary goals of measuring uncertainty is to use it to compare different data sets to determine if they are consistent. Without access to more advanced statistics, this becomes

a question of “agreement”, not consistency: is it more likely than not that the data sets were drawn from the same broader data pool? If the data ranges defined by the uncertainties are very far apart, then the data sets do not agree; if they overlap strongly, then the data sets do agree. In between, the answer is unclear: this is the realm of careful statistics.

Data range consistency has been well-studied, especially by Kung. In her comprehensive lab system (discussed more fully in Chapter 2), uncertainties were carefully defined via standard deviations; in her survey questions, students chose not only whether the data agreed, but why they thought so [36]. The rubric for their responses is shown in Figure 4-8.

Even though Kung’s survey questions were designed for a more statistically oriented class, they were appropriate for 1-Series assessment as well. While the calculation and meaning of the standard deviation is different from the upper to lower limit uncertainty used in the 1-Series, both tools are used for the same overall purpose: evaluating the quality (especially precision) of experimental results, and, most importantly, allowing the comparison of two data sets (or one data set and a reference value). Indeed, even with their further statistical training, Kung found “the difficulty is not with the calculations, but with the knowledge that, and how, uncertainty concepts should be applied to all measurements” [23].

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**

- 18% 1. There isn't a significant difference between the two group's results.
- 16% 2. Everything has error, it's impossible to get exactly the same every time.
- 7% 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.
- 21% 4. Group A's range is from 429 to 439, group B's from 436 to 448, so the ranges overlap.
- 1% 5. Other

**No**

- 12% 6. Their averages are different.
- 20% 7. There is a significant difference between the two group's results.
- 7% 8. The difference of eight millimeters is a large difference compared to the distances they're measuring.
- 1% 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.
- 11% 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.
- 8% 11. Group A's average of 434 does not fall within the range for group B (436 to 448), and vice versa.
- 2% 12. Other

**Figure 4-8: Range measurement question and results from Kung's Scientific Community Labs (reproduced from [36]).**

A modified version of Kung's classification scheme was used for the 1-Series survey items. Instead of students choosing one of the arguments in Figure 4-8 from a list, they wrote out their arguments in text, which was subsequently classified via Kung's rubric. Why make this change? Kung's online multiple choice-based survey system required students to select whether they thought the two data sets agreed or disagreed first, and then chose reason(s) for that decision; this prevented students from choosing arguments that seemed convincing from both the "yes" and "no" sides simultaneously. A short answer-based classification system allows for students to choose both "yes" and "no" answers if they feel so inclined; one might expect students to work through arguments from both sides before making their decision. Moreover, it allows students that feel the data is truly ambiguous to communicate that uncertainty. This was especially important considering the uncertainty pedagogy used in 1-Series only yields conclusive answers for some kinds of data. The modified rubric used is shown in Figure 4-9.

Two sets of these Kung-style "agreement" questions were asked in Fall 2012 in 1AL, one pre-instruction and the other post-instruction. Following her treatment, the two data sets were designed to have different averages and uncertainties, as non-experts tend to focus on these features rather than the overlap of spreads; the prompts for these are shown in Figure 4-10. The data for each rubric response for both surveys are shown in Appendix V.

For the pre-instruction survey, the data ranges do overlap, although the averages are not included in each others' uncertainties, which is a signifier of very strong overlap. Still, the expectation based on the 1-Series pedagogy would be that the data sets did agree, as the range overlap is substantial.

**YES:**

- A) There isn't a significant difference between the group's results
- B) Everything has error, it's impossible to get exactly the same every time
- C) There's a difference of XX between the two group's averages. XX is small, so they pretty much have the same result
- D) Group A's range is from XX-YY, B's from ZZ-AA, so the ranges overlap
- E) There is a possible value that could fit both ranges
- F) Other yes

**NO:**

- G) Their averages are different
- H) There is a significant difference between the two group's results
- I) The difference of X is a large difference compared to the times they're measuring
- J) Group A's range has a width of XX, B's has a width of YY, so they have different results
- K) Group A's range is from XX to YY, B's from ZZ-AA; they only overlap for CC which is not enough
- L) Group A's average of XX doesn't fall within the range for group B (YY-ZZ) and vice versa
- M) There is not a possible value that could fit both ranges
- N) Other no

**NOT SURE**

- O) Group A's range is from XX to YY, B's from ZZ-AA; they only overlap for CC, which I am not sure is enough overlap for them to agree
- P) There is a possible value that could fit both ranges (as in E, but not sure whether this means they agree)
- Q) Other not sure

**Figure 4-9: General rubric used for range comparison questions.**

Generally, students believed the data sets agreed: 60% of arguments held that the data were in agreement, as opposed to 26% disagreement and 13% not sure. Students generally found that

the data was clear cut: only 3 students out of 45 provided both an “agree” and “not sure” answer; no student offered an argument that the data agreed and also an argument that the data disagreed.

Around half of the arguments made were classified as clear spread arguments (categories d, k, and o, from the rubric in Figure 4-9). Of these, 67% argued that the data sets agreed, 22% weren’t sure whether or not they agreed, and only 7% argued the data sets disagreed. 40% of arguments overall held that the data agreed because the spreads overlapped.

Non-spread arguments were largely scattered evenly around the other classification categories. Two were more noteworthy: category “e” and category “l”. Category “e” examines both data sets to try to find a common value that could fit both data sets, and bases agreement on whether or not there is such a value. For example, in the case of the problem shown in the pre-instruction survey, a  $t=735$  ms value lies within both ranges. 13% of students believed the data agreed for this reason. This is a pernicious line of thought, since it matches up well with the “any overlap is agreement” approach, which is a spread-based approach. Though this reasoning is adjacent to spread reasoning, it is still hunting for a true value, in this case, the value that allows both data sets to be consistent with each other. Not only is there no true value for the data, but also even if there was, nothing necessitates it being the convenient middle value anyway. Category “l”, selected by 8% of students, is the aforementioned statement that the data can’t agree because the data sets do not include each other’s average value; although this is a reasonable rule of thumb for strong agreement, it is not itself proof of disagreement if the criteria is not satisfied.

After instruction, students were asked a similar question, with the main difference being only a glancing overlap between the two data sets, where only a single value (83 cm) was present in both data sets.



6. Two groups of students measure the time it takes for a 100 g ball to fall 2.5 meters. Group A obtains the result  $t = 732 \pm 6$  ms, while Group B measures  $t = 741 \pm 7$  ms.

Do the results of the two groups agree? Why / why not?

Yes	No	Not sure
32	14	7
71.1%	31.1%	15.6%

6. Two groups of students measure the range of a ball that has rolled off a ramp. Group A measures  $x = 86 \pm 3$  cm, while Group B measures  $x = 81 \pm 2$  cm.

Do the results of the two groups agree? Why / why not?

Yes	No	Not sure
141	43	24
89.9%	27.4%	15.3%

**Figure 4-10: Percentage of students making each type of argument. (Totals greater than 100% because some students made multiple arguments.) Pre-instruction (above, N=53) and post-instruction (below, N=208) questions and data shown.**

Despite the more tenuous overlap of the second data set, 77% of students made at least one argument supporting agreement, as opposed to 62% before. This difference is significant at the  $p=0.02$  level. Because students were able to make arguments both supporting and denying agreement, though, this does not mean students were more confident in the agreement of the second data set. Only 55% of students made only agreement arguments for the pre-instruction data set; this only rose to 61% for the post-instruction data set, less than the increase of total agreement arguments overall.

After instruction, students were more likely to make multiple arguments, including arguments that lead to different conclusions. In the pre-instruction sample, only 18% of students made more than one argument to support their assessment; in the post-instruction sample, this rose to 28%. Although this was not a statistically significant change ( $p=0.08$ ) on its own, it does seem to be a major contributing factor in the increase of “yes” answers mentioned above.

Even though 77% of students included an argument in favor of agreement, 21% of these also included an argument supporting disagreement or that the relationship between the ranges was unclear. This was true of only 10% of such students pre-instruction. In total, 25% of students either chose an “unsure” argument, or presented arguments from both “yes” and “no” agreement sides; this was only true of 15% of students in the first survey. The percentage of students making only “no agreement” arguments dropped from 29% to 14%, despite less agreement between the two data sets ( $p=0.01$ ). This increased mixing of “no”, “yes”, and “not sure” arguments is a positive, since the agreement of the two data sets is certainly ambiguous.

In some cases, the student does not necessarily indicate a final decision as to whether or not the data sets agree: “the best estimates of the two groups do NOT agree but the lower estimate of group A and the lower estimate over group B do match.” In other cases, students do draw a conclusion, but it is tentative because opposing arguments are persuasive also: “barely; there is a slight overlap of measurements (83 cm) but for the most part the groups don’t agree.”

The increased frequency of “yes” arguments (concerning the more tenuously overlapping ranges) were not due to decreased use of spread reasoning by the students. To the contrary, the number of students making at least one spread-based argument grew significantly, from 51% ( $\pm 7\%$ ) pre-instruction to 65% ( $\pm 4\%$ ) post instruction ( $p=0.05$ ). If more students used spread

reasoning, why don't more of the arguments oppose the agreement of the data sets, since the overlap between them is so tenuous?

Pre-instruction, there was a significant ( $p=0.02$ ) correlation of 0.34 between students choosing "yes" answers and students using spread reasoning. Moreover, there was a stronger anticorrelation ( $p=0.001$ ) of -0.45 between students choosing "no" answers and students using spread reasoning. Essentially, the overlap in the pre-instruction survey was obvious enough that the main way to think that the two data sets did not agree is if one didn't use spread reasoning at all. Some of the students pre-instruction were already able to use range overlap to determine agreement, and they overwhelmingly found that the data sets agreed.

Post-instruction, all of these correlations vanish. First, more students were using spread reasoning, so spread reasoning should have less correlation with any particular answer. Second, the actual problem was more ambiguous; since the overlap was more tenuous, spread reasoning could plausibly be used to support any of the three cases for agreement / disagreement / not sure. (Interestingly, the only correlation between spread and agreement type to even approach significance was the "not sure" option, with a  $p$ -value of 0.10)

While spread-based reasoning was more common with the post-instruction data set, so were some more point-like comparisons. Some students (rubric categories e&m) focus on whether there was a possible value that could fit both data sets and then use that to determine whether the data sets agree. These may be transitional students: although the excessive focus on particular numerical values (in this case, 83 cm) is not really spread-like, those values can only be found by examining the ranges of each data set. 18% of arguments in the pre-instruction survey used this argument, compared to 41% post-instruction. The form of the problem probably primes these sorts

of arguments; since the data sets overlap only at one particular value, it draws attention to that number. In the case of the second problem, the lower limit of range one and the upper limit of range two land on the same value, 83 cm. 87% of students using such an argument thought that the two data ranges agreed, despite the two ranges barely meeting; because so many of those that choose this option seem to believe it denotes agreement, additional emphasis on that argument could have caused the increase in “yes” answers seen on the second survey question.

Ironically, then, a more tenuous overlap may draw focus to the overlapping point, priming more students to argue that the data sets agree. This response is encouraged by our uncertainty technique’s focus on calculating and propagating the error through edge points of the range; by giving these edge points privileged status in the calculations.

### **Interpretation of uncertainty**

Students were expected to summarize their data as average  $\pm$  uncertainty (see Chapter 2 for more discussion about these expectations). Even though the uncertainty as used in 1-Series labs does not have the technical meaning of a standard deviation, it is derived from and represents the range of possible values that students recorded or calculated in their measurement of that quantity. In one survey question, students were asked to state their own understanding of what the “average  $\pm$  uncertainty” means about the measured quantity. The following question was posed to students in Winter 2013 for 1BL, pre-instruction:

**“A labmate measures the velocity of a glider as  $1.37 \pm 0.08$  m/s. What does that imply about the velocity of the glider?”**

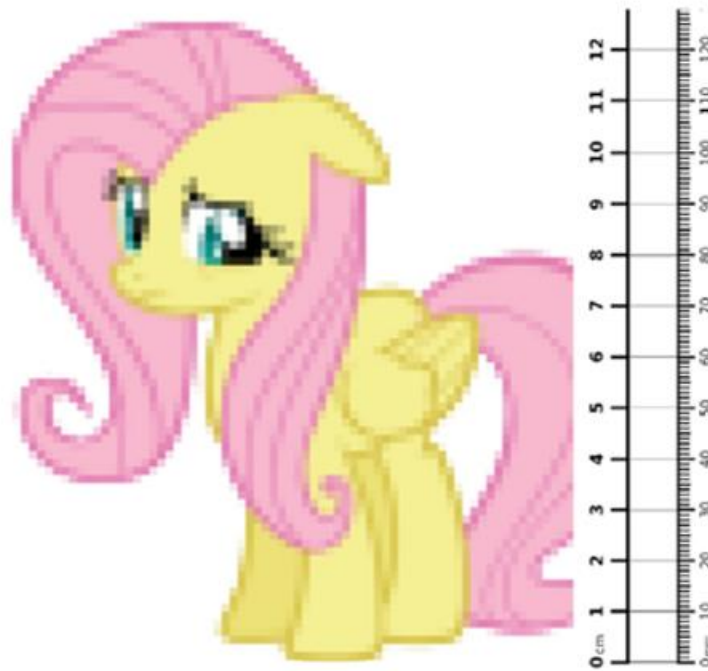
Most of the students in that class (69% according to their survey responses) had previously completed a semester of the new curriculum, whereas the remainder had not been exposed to that formulation of expressing range at least in UCSD physics. Both groups would have been familiar with 1-Series labs measuring the velocity of a glider on an air track.

Of those who responded, 49/78 students (63%) calculated the lower and upper limits of the range and stated the velocity was within that range. 14 of the remaining students (18%) made more general statements that the velocity of the glider had different possible values. This is a basic operational understanding of the notation and its meaning: that the value of the velocity was within that range.

**Pixel size: uncertainty from properties of the object.**

In Spring 2013, the third and final physics class in the sequence (1CL) began using the new changes to the lab curriculum. This was the third “on-sequence” lab class, and the first opportunity for students to have their entire 1-Series lab experience be within the new environment. Indeed, of the 334 students answering pre-instruction surveys, 234 had taken the modified 1AL in the Fall (72%), and 258 (79%) had taken the modified 1BL in the Winter; 230 (69%) of them had taken both classes in a row.

In the pre-instruction survey, students were presented with a modified version of an arrow measuring exercise, with a focus on measuring, estimating uncertainty, and comparing to the (hypothetical) results of classmates. This allowed the testing of both uncertainty skills and how students communicated with lab groupmates that may have had different results.



2. The above picture is pixelated. Using the ruler on the right, to what precision could you determine its height? (Assume you would be measuring from the bottom foot to the top of the hair.)

3. A lab-mate reports the height as  $124 \text{ mm} \pm 1 \text{ mm}$ . Would you agree or disagree with this statement? Why or why not?

**Figure 4-11: Measuring and communicating uncertainty for a pixelated image.**

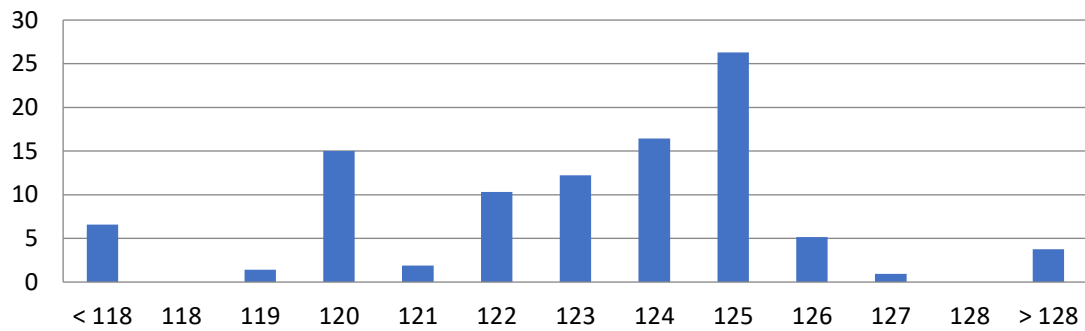
As shown in Figure 4-11, the image is noticeably pixelated and flanked with a ruler. In contrast to the pilot study's arrow measuring problem (where the ruler markings were too far apart, leading to lower than expected measurement uncertainty based on the ruler), now the object itself is the source of the measurement uncertainty. Because of the 2 mm pixelation, the ruler increment uncertainty is too small to represent the uncertainty in the height of the object. For the eagle-eyed, the foot of the picture was not lined up precisely with the bottom of the ruler. Both of these affect the comparison with the lab-mate's height comparison of  $124 \pm 1 \text{ mm}$ ; the 124 mm height was

chosen to match the head height of the picture and ignore the slight misalignment of the feet with the end of the ruler.

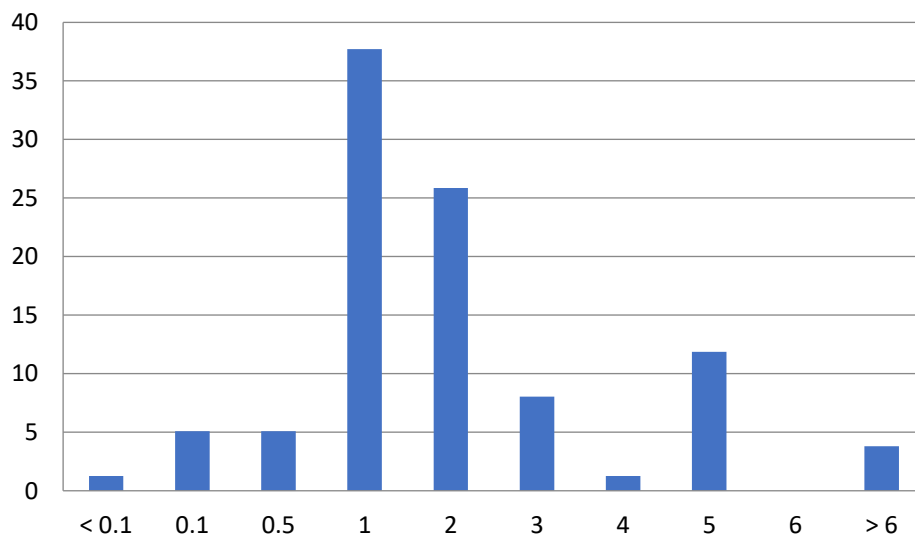
Although the question did not explicitly ask for a height measurement to be recorded, 219 students provided one regardless; these heights are shown in the histogram in Graph 4-12. Because the ruler had both millimeter and centimeter markings, students could express their heights and uncertainties in either units system. Height and uncertainty measurements have been converted to mm for ease of comparison. Roughly half of the students used each measure of height. Mostly, students' own reporting of the measuring unit of the height and uncertainty were used; but in cases where students clearly seem to have made a decimal place error (e.g., recording the height as 120 cm), the correct measuring unit has been substituted.

Graph 4-13 shows student determinations of the uncertainty. If students did not consider the effect of the pixelization on the length uncertainty, we would expect them to choose an uncertainty of  $\pm 1$  mm, or, perhaps, half that number. Unlike in the case of the arrow measurement, this does not represent an inappropriate reliance on the minimum increment of the ruler; because the ruler marking on the mm ruler are so small (and the exact line-up with respect to the ruler would be difficult to determine), 1 mm would be an appropriate estimate with no pixelization. Indeed, 38% of respondents chose that value, with only 11% of students choosing a smaller value of uncertainty. The real number of students that feel the uncertainty is smaller than the 1 mm increment is likely smaller, since the very small values include some apparent calculation mistakes.

**Graph 4-12: Percentage histogram of height estimates (N=219).**



**Graph 4-13: Percentage histogram of uncertainty estimates (N=236).**



Still, the plurality of responses (47%) chose an uncertainty between 2 and 5 mm, inclusive. These larger values seem to be reflective of the effect of the pixelation on the uncertainties involved, as the pixelization itself is 2 mm tall.

In addition to reporting the best estimate of the height and the uncertainty of that number, students were asked to decide whether their estimate agreed with that of a (hypothetical) lab-mate,



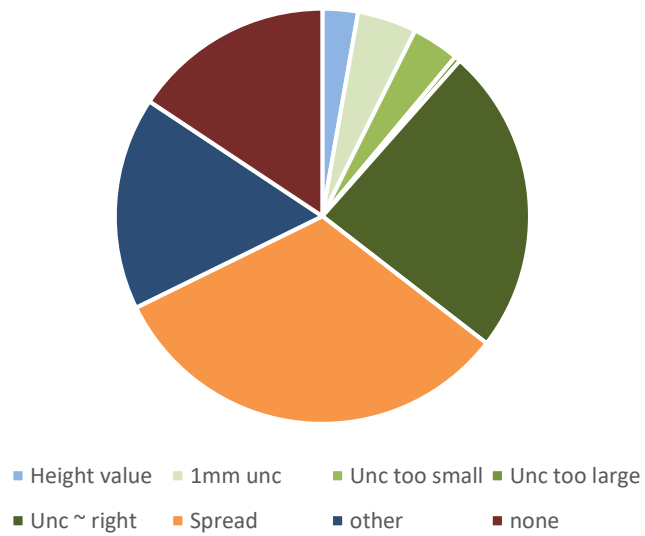
who reports a value of  $124 \pm 1$  mm. Although the average of this is reasonable, the uncertainty does not include the effect of the pixelization.

What reasons did students provide for agreeing or disagreeing with the lab-mate's estimate? Most responses either addressed the height estimate alone; the uncertainty estimate alone; or used a spread method to compare the lab-mate's range with the student's own estimate. Overall, 59% agreed with the lab-mate's estimate (59%), and 35% did not. The reasons the students used did correlate with the conclusion the student reached, as can be seen by comparing the Graph 4-14 (for agreeing students) and Graph 4-15 (for disagreeing students).

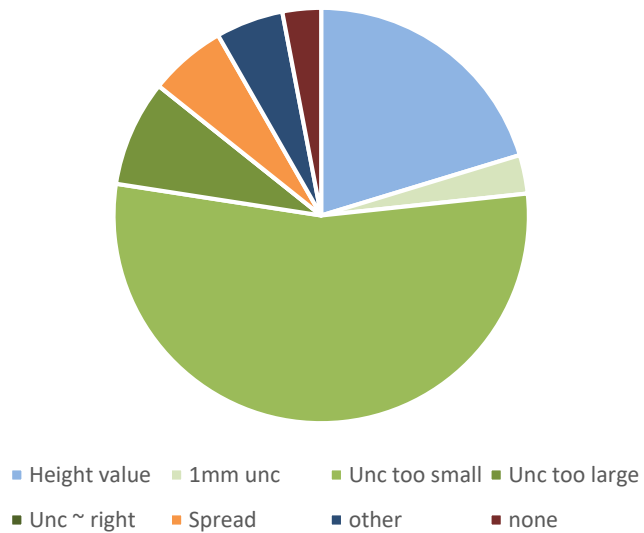
For those who agreed, the most students (35%) used a spread approach to determine that their range overlapped with the lab-mate's range. 26% of agreeing students mentioned that the lab-mate's uncertainty estimate was correct. These students either did not consider the effect of the pixelization or perhaps underestimated its magnitude. 17% of students provided no explanation at all for their conclusion. Some students who ultimately agreed with the lab-mate's range still thought that the uncertainty of 1 mm was too small, but still felt the ranges agreed.

By contrast, students who disagreed with the lab-mate's range estimate mostly considered the height or uncertainty to be incorrect in isolation. 62% described the 1 mm uncertainty as too small, and 23% didn't agree with the height estimate. Only 8 students (7%) used spread reasoning to discount the lab-mate's range estimate; as will be seen, this is because most disagreeers still found some overlap between the two range estimates. Interestingly, only 3% of disagreeers failed to provide a reason for the disagreement, in contrast with 17% of agreeers.

**Graph 4-14: Justifications for students who felt their estimate agreed with the provided average and uncertainty estimate. (N=198).**



**Graph 4-15: Justifications for students who felt their estimate disagreed with the provided average and uncertainty estimate. (N=117).**



The difference in the amount of spread reasoning can be investigated by comparing the lab-mate's range to that of the students (for the students that provided a full uncertainty range). There were 141 students that provided a length, uncertainty, and an answer explaining their reasoning for agreement or disagreement. Each student range was compared to the 123-125 mm range of the lab-mate; if the student range overlapped with this range at all, they were classified as agreeing with that range. (Although this criteria is not what we applied when teaching range comparisons, as it excludes the crucial "neither agree nor disagree category", this maps closer to our typical students' understanding of spread measuring, as can be seen above in "Comparing two ranges")

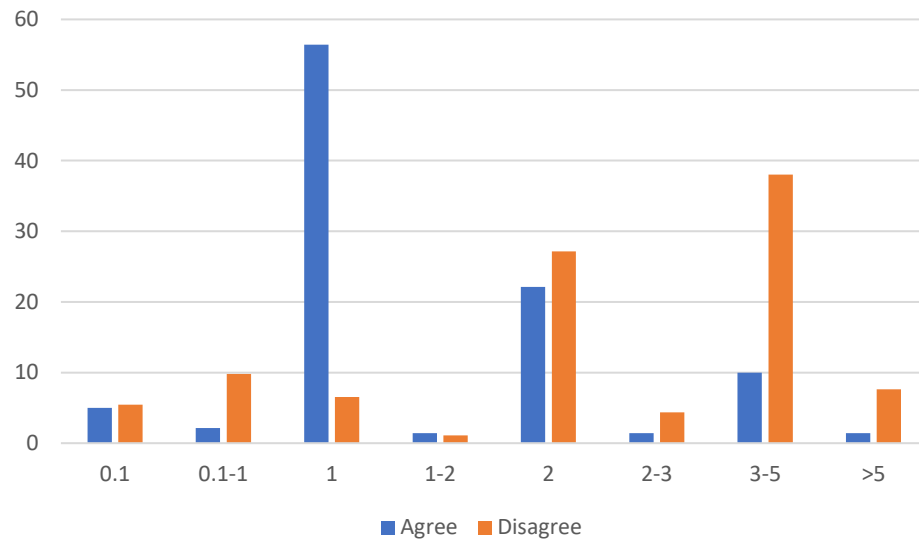
Of these students, 84% had ranges that overlapped with the lab-mate's. Did the students who reported a range correctly use spread arguments more frequently than those that did not report their range? In total, 198 students believed that their data agreed with their lab-mate's data; of these, 97 students also reported a range. 41% of these students use spread as their reasoning for the agreement, as opposed to only 30% of students who did not report a full spread (or best estimate  $\pm$  uncertainty). This is a significant difference at  $p=0.05$ . It is possible that writing down the best estimate and uncertainty primed them for the type of range comparison with the lab-mate's range, since that is a typical lab task; alternately, students that used range as their justification may have been more likely to write down that range as a way of showing their work.

This difference is only found for students who thought the ranges agreed. Of the 45 students who disagreed with the lab-mate's estimate, only 3 (7%) used spread reasoning to justify their disagreement. These mirror the overall numbers for disagreeing students as shown in Graph 4-15 (also 7%). In part, this is because a majority (32/45) of disagreeing students had ranges that overlapped with the lab-mate's range; consequently, they couldn't use range reasoning to explain

their disagreement. Still, of the 13 disagreeers whose range did *not* overlap with the labmate's, only 3 used spread reasoning to justify their disagreement (23%). This is about half of the rate agreeers used spread to justify agreement. Even though the statistics for the disagreeers are small and prevent any statistical conclusion, this does make some sense. Any overlap between two ranges can feel like a conclusive agree; but non-overlap may not feel like a conclusive disagreement. This asymmetry may be further exacerbated because the problem frames the comparison as being with a lab-mate, with whom one needs to build consensus as part of a lab group. One student agreed with the measurement despite having a 3 mm uncertainty of their own, because the hypothetical lab-mate "might have more accurate vision than I." Another student considered it "a reasonable estimation, but since it is hard to tell the height, a wider uncertainty would make more sense." Rather than using the class's technique to compare the range's mathematically, the students are considering whether their peer's range is acceptable, even if it is not really correct.

Still, these lab frame concerns are unlikely to be the dominant issue at stake here. This is because disagreeers are most likely to disagree because the lab-mate's uncertainty is too small; 62% of disagreeers used that reasoning. This implies that those disagreeers have recognized the pixelation, and chosen an uncertainty value higher than the 1 mm value of the lab-mate. But choosing this larger value of uncertainty, ironically, only increases the chance of spread-based agreement, because the student's own data range is more likely to overlap with the lab-mate's because of their own larger spread.

**Graph 4-16: Percentage histogram of student uncertainty estimates, based on whether they felt their range agreed or disagreed with the provided range of  $124 \pm 1$  mm.**



Graph 4-16 compares the uncertainty estimates of agreeers and disagreeers. There is a moderate anticorrelation ( $r = -0.31$ ,  $p < 0.001$ ) between a student's uncertainty and agreeing; that is, students that agreed chose smaller uncertainties. These correlation is mostly found in the 1 mm uncertainties (dominated by agreeers) and the 3-5 mm uncertainties (inclusive, dominated by disagreeers). 1 mm uncertainty estimators chose agree on a 79-6 basis. While 2 mm uncertainty estimates were fairly common for both groups, estimates higher than 2 mm were predominantly disagreeers at a rate of nearly 3-1 (42-16).

The relative dearth in disagreeers using spread reasoning makes sense, then: 76% of them chose large values of uncertainty equal to or greater than 2 mm, making their ranges very likely to overlap with the spread of the lab-mate's range. If they used spread correctly as a part of their reasoning, most would have concluded that the two data sets agreed. Instead, they decided that different values of uncertainty (or best estimate) were too different to accept the lab-mate's

estimate. This is appropriate from a lab frame perspective: if your lab partner determines or estimates an uncertainty that misses a critical systematic, that should be addressed before using a range overlap approach to establish agreement.

This tension between spread reasoning and different determinations of uncertainty is most apparent among the agreeers that estimated values of uncertainties of 2 mm or greater: they have chosen a larger value of uncertainty than the lab-mate's estimate, but still agreed with their overall estimate anyway. Of agreeers with uncertainties greater than or equal to 2 mm, 45% used spread reasoning; spread reasoners made up only 31% of agreeers with smaller uncertainties (significant at  $p=0.05$ ). Although this makes sense, what reasons did students who chose large uncertainties use to justify agreement?

These students mostly replied with brief statements of general agreement despite the differences in the uncertainty. One student, whose estimate was  $12\text{ cm} \pm 5\text{mm}$ , wrote only “agree, because it's the same as mine”, seemingly unperturbed by the large difference in uncertainties. In some cases, this seemed to bleed into agreement in the sense of agreeing with a lab partner who might have a slightly different idea: different, but not enough to really matter. “Agree, they got the same measurement I did and maybe they had a more accurate way of measuring to have lower uncertainty.” One gets the sense these distinctions might have been ironed out in a group discussion in an actual lab situation. This basic question – how do lab students build consensus about uncertainty? – might be studied by tracking student conversations in the same style as Kung and Linder [37]. In that study, they recorded student conversations and measured what prompted sense-making discussions as opposed to logistical talk. See Chapter 5 for further ideas of how to measure how lab groups collectively determine uncertainties.

## **Acknowledgements**

Chapter 4's pilot study is material being prepared for publication as "Implementing algebra-based measurement uncertainty techniques for non-majors". Schanning, Ian L; Anderson, Michael G. The dissertation author was the primary investigator and author of this paper.

## **Chapter 5**

### **Chapter 5 : Responsibilities to biology students**

Ultimately, the goal of this project is not just to make a specific change in the curriculum of an introductory physics class, but to build on that change to continue to find better ways to help biology students learn about measurement and uncertainty. The context of a service class like the 1-Series makes this more complicated, however, since we are not “building the next generation of physicists”, but rather trying to contribute to support the next generation of biologists! While practiced physics educators are comfortable with physics content as taught to biologists, that doesn’t mean we should be able to decide on our own which subjects are taught and which are not, especially in regard to uncertainty. This is even more complicated for the biology students, who have to juggle the norms of their own subject, physics and chemistry in lower division classes. With a look towards the future, this chapter addresses the following questions:

- What should the role of physics in laboratory education for biology majors be?
- How did students perceive the roles of physics and chemistry lab instruction?
- How can the lessons of this project be applied to future iterations?

#### **Biologists in physics labs**

To some extent, determining the goals of uncertainty instruction for physics majors is straightforward. The instructors and graduate students are physicists that have more experience



than their own students, and so they can examine their own experiences to determine what concepts and skills they find to be the most important for a physicist. For example, Deardorff [8] simply asked a population of graduate students and professors: “What do you think are the most important concepts or skills students should learn about measurement uncertainty and error analysis?” The categories resulting from answers to this question were reasonable and expected for physicists:

- All measured values have uncertainty
- Uncertainties must be estimated and clearly reported
- Reporting the proper number of significant figures
- Propagation of errors
- Identify and classify sources of error
- Interpreting and reducing errors
- Use of uncertainty for comparing results or designing experiments

Similarly, many physics departments have written and publicized learning goals for the major that go beyond the physics concepts typically covered in classes. UC Berkeley’s physics department, for example, emphasizes project-building, communication, safety, and the ability to pursue career objectives in addition to knowledge of physics topics [38].

These goals may not be appropriate for all physics undergraduates, since they are written by faculty members that have primarily followed an academic and (at Berkeley) at an R1 level in physics. This represents the minority of physics graduates; less than 40% move into graduate study in physics or astronomy a year after graduating, a finding that has been consistent of more than a decade [39]. Still, they are at least informed by the experience of experts working in plausible future occupations for many physics undergraduates.


When considering the instruction of non-major students, particularly biology majors, the determination of what should be the focus of instruction becomes significantly more complicated. Biology majors have required science courses in two other outside departments – Chemistry and Physics – and often are required to take those courses, lecture and lab, before coursework in their major. One could take the tack that it is solely the responsibility of biology departments themselves to educate their majors on scientific literacy and measurement (in addition to biology content), and leave the instruction of basic physics and chemistry to their respective departments. There are some problems with this approach, however, that can be more easily seen by examining, as a case study, the biology department at UCSD.

Learning goals for majors are available from the departmental website, and differ depending on the particular biology topic area; for simplicity, only the general biology goals will be considered here. Omitting the specific topic-based learning goals exclusive to biology, students are expected to [40]:

- Understand how mathematics, physics and chemistry are integrated into the study of biology
- Construct reasonable hypotheses to explain biological phenomena and design effective experiments to test the hypothesis
- Implement contemporary biological research techniques to conduct experiments, and use quantitative and/or statistical approaches to analyze the results and draw appropriate conclusions from them
- Use digital technologies to search the scientific literature, and to retrieve and analyze information from reliable databases.

Taking the conservative approach to instruction of biology majors, then, would mean that physics and chemistry professors would only focus on the first of these goals. The other three goals, however, are experimental in nature: hypothesis construction, experimental design, analysis, drawing conclusions, and accessing previous work. These map directly to the scientific method, and therefore have a natural place in a laboratory class as opposed to the lecture part of the course.

Why not let the biology lab classes handle the laboratory science and general science learning objectives? While this might be an option at some schools, at UCSD it is made more difficult by the way the biology major is set up, as seen in Figure 5-1.



**Biological Sciences**  
where discovery comes to life  
UC San Diego

# General Biology (BI31)

## SAMPLE PLAN

Biology Undergraduate Student and Instructional Services | 1128 Pacific Hall | (858) 534-0557 | [www.biology.ucsd.edu](http://www.biology.ucsd.edu)

	FALL	WINTER	SPRING
FR	CHEM 6A MATH 10A	BILD 1 CHEM 6B MATH 10B	CHEM 6C CHEM 7L MATH 11* BILD 3
SO	BILD 2 CHEM 140A PHYS 1A&1AL	BICD 100 CHEM 140B PHYS 1B&1BL	UD Bio Lab PHYS 1C&1CL
JR	BIBC 102 UD Bio Elective	UD Bio Elective UD Bio Elective	UD Bio Elective UD Bio Lab
SR	UD Bio Elective	UD Bio Elective	UD Bio Elective

**Figure 5-1: General Biology degree plan at UCSD [41].**

Biology undergraduates' most likely progressions can be approximated by using the sample degree plan provided by the department; lab classes are labeled with an L. Only two of the six required lab classes are biology classes—three belong to the introductory physics series, and the sixth is the notorious chemistry lab, Chem 7L. Not only are the biology lab courses deemphasized in terms of number, but they are also placed relatively late in the degree; a student on the general biology default plan will have taken three physics labs and one chemistry lab by the time she finishes her first biology lab class. Such a student will have practical scientific habits built up from her several quarters in chemistry and physics lab classes, regardless if those classes address the issues directly or not. It is important, then, for both chemistry and physics lab classes to help biology majors achieve their departments' scientific literacy and method learning goals. It may be the students' first exposure to taking their own data and drawing conclusions from it, so it is important that these laboratory classes focus on these goals in addition to teaching their technical content.

Still, the teaching of lab skills in non-major classes feels like a stopgap solution: why does the biology major lack an early biology lab class? At UCSD, this is partly because of how chemistry classes for biologists are structured. Unlike in physics classes (where the lab and lecture classes are corequisites), the first two chemistry classes (6A/6B) are prerequisites for the required chemistry lab (7L), which is in itself a prerequisite for the upper division biology labs. As will be seen in the subsequent discussion of chemistry labs, these upper division biology labs expect the chemistry lab class to teach basic safety, equipment, and organizational skills that they rely on in their upper division work. While this all makes a sort of sense, biologists should be ultimately

responsible for the teaching of best lab practices for their field, which implies the development of a lower division biology lab class to start the process early<sup>11</sup>.

In the meantime, chemistry and physics lab classes have the responsibility of teaching future biologists about the practice of science. While this responsibility is diffused over the three separate lab courses in the case of physics, it is concentrated in one single taxing course in chemistry: Chem 7L. Coincidentally, this course had recently also been revamped by the chemistry lab instructor Dr Sandrine Bernoilles, who I interviewed about why and how she redesigned the class. Comparing the design goals and implementation of the two lab courses can show how well they combine to cover the full practice of science for their biology students.

### **Lab redesigns in physics and chemistry.**

Before discussing how chemistry lab instruction differed, it makes sense to repeat the top-down instructional goals for uncertainty for the physics lab classes. In short:

- Calculate or estimate the measurement uncertainty for *each* measured quantity
- Interpret this uncertainty as a measure of variability of the quantity
- Judge the consistency of multiple data sets (or a single data set with a theoretical quantity) using this variability
- Propagate uncertainties to facilitate meaningful comparisons, if necessary

---

<sup>11</sup> As this dissertation was being completed, the biology major did in fact implement such an undergraduate biology lab class.

Although these goals summarize the uncertainty part of the class, the following broader goals were also implicit in course design:

- Precision rather than accuracy is emphasized, with most conclusion questions consisting of a comparison of different measurements rather than comparison to some objective correct answer
- Little focus on correct lab procedure, because of the deemphasis on accuracy of results.
- Little expectation for students to know base physics as they begin the lab classes, because of the co-requisite status of the class

These goals informed the way lab tasks were designed. The lab experiments focused on demonstrating important conceptual ideas (ie, Newton's 3<sup>rd</sup> Law) rather than measuring important quantities. Although there was some continuity from lab to lab in terms of equipment (especially via the LoggerPro data acquisition system), learning how to use the equipment was more a means than an end. There was no expectation that students would use sort of equipment in their future post-lab experience as non-majors, so detailed investigations into the experimental equipment sometimes seen in physics major classes were omitted. In the typical physics undergraduate experience, by contrast, significant time is spent learning how to operate calipers, oscilloscopes, and op-amps, with the hope that those skills would transfer to later research projects.

Moreover, the goals were built into the primary assessment measure of the class: the lab report. As was discussed in Chapter 2, the report consisted of conceptual questions, data reporting, and analysis of that data using measured and propagated uncertainties. We chose to focus on a question and answer format (rather than a formal lab report) to enable us to ask detailed questions about uncertainty with escalating sophistication as the class proceeded. This extended to the point

that even common lab report features (like a description of the lab procedure) were omitted entirely.

Similarly, the lab's singular focus on measurement and uncertainty was built into both the conclusion questions and the in-class checkpoint questions that often centered on gathering measurement uncertainties during the class. One ideal effect of all this repetition is to encourage a lab frame that centers the recording of measurement uncertainties. For the vast majority of labs, this meant student procedural errors could have little effect on the final grade: as long as one made the proper analysis of what one's data did say, then how close that result came to the perfect version of that experiment was irrelevant.

Sometimes, students attended office hours because they perceived that their experiment had gone wrong, and they were hoping for some triage from the LTAC on duty. In the worst case, students came to office hours having forgotten to record the required uncertainties in their measurements that they needed to answer the questions for their lab reports. In those cases, we advised the students to recall the experiment, and estimate a value for the relevant uncertainties (and provide a note to the TA, just in case their memories were inaccurate!). To a fault, then, the focus was on "what does your data say?" rather than producing data that achieved a certain expected result.

By contrast, Dr Bernoilles [42] redesigned the chemistry lab class so that students should be able to:

- Recognize potentially dangerous chemicals and handle them in a safe way
- Record measurements and notes in an organized lab notebook
- Perform common experimental procedures (like titration) to a certain accuracy
- Write formal, research-like lab reports, including background chemistry that was not taught in the lab course itself.

These goals point to a different emphasis than the 1-Series physics lab class. This is partially due to differences in the fields of study, but mostly because of chemistry's different role in supporting the biology curriculum.

Introductory chemistry labs focus on procedure and equipment because, generally speaking, the same equipment and techniques will be used in subsequent biology labs. These basic techniques (like proper use of pipettes to transfer fluids) will be used repeatedly in those later lab classes and future research, and the biology departments expect those skills to be taught via chemistry labs, at least initially. This is similar to the way that introductory physics labs for majors will often teach the use of circuits, assuming that will pay off for subsequent experimental researchers.

While physics labs may have some nominal constant safety rules (no open-toed shoes) or more important guidelines in the case of one-off equipment like lasers, chemistry and biology lab spaces are more dangerous, even at the introductory level. Toxic chemicals are transported by students in breakable glass, which can itself cause injury. Needles and biohazards need to be disposed of in specific ways. Unlike in physics labs (where in my experience students asking about lab coats has been something of a running joke), there is the possibility of spilled concentrated acids or bases being absorbed into clothing, potentially leading to burns. All of these factors necessitate a serious approach to teaching lab safety, which is why it remains a priority for the chemistry lab courses. In the Chem 7L class, students were required to briefly research the chemicals used in each lab, which would help to alert them to the potential hazards well in advance. These facts were assessed by pre-lab quizzes for credit, which encouraged students to read ahead and be mentally prepared for the safety procedures that would be required for each lab. This not



only helped students be more safe in lab, but also provided a model for how to approach safety concerns responsibly in subsequent research work.

This more organized approach also applies to how chemistry students are instructed to take lab data. In physics, lab notes are something of an afterthought; data values are mostly inserted into pre-written tables in the lab manuals, with the occasional handwritten notes taken in the margin. Although we instituted an informal check of lab data before students left class (to ensure that they had the appropriate data and uncertainty values to help answer the lab conclusion questions), the actual lab data itself was graded in a perfunctory way. As long as students had recorded lab data that seemed vaguely plausible, they received full credit.

Chemistry 7L took a stricter tack, necessitating the use of student-written lab notebooks to document all of their observations during the lab. Teaching assistants enforced the key rule: only notes that were taken during the lab, in the notebook, were allowed to be referenced in the formal lab write-up. This was intended to support a lab frame where students would take detailed and careful notes on their experiments, emulating the attention of a researcher working on their own research project. Because the teaching assistants would occasionally examine the state of the lab notebook, student notes needed to be clear and readable to a different person besides themselves. These are excellent practices for a research lab, where later graduate researchers often consult the experiments and writing of the previous generation of student researchers and postdocs.

With those more detailed notes, students were able to complete a more comprehensive and professional lab write-up, written intentionally to emulate a research paper. The lab write-ups were typically several pages long, consisting of hypothesis, background, and analysis sections. These more elaborate lab reports required more time to complete; consequently, a new lab was performed

only every other week, with the remaining extra time allotted to completing the report. Because the general chemistry lecture class (Chem 6A) was a prerequisite for the lab class, students had familiarity with the theory behind the labs. This enabled them to write a more confident and correct conceptual treatment in their reports than would have been possible for the physics students, who at times were learning the concepts for the first time in their 1-Series lab class.

Finally, the additional lab time allowed for more complex lab procedures. The procedural focus meant that a certain minimal level of accuracy was required for full credit for the lab. This might mean a certain amount of a compound should be crystalized, or a certain precision was required for acid-base titration. Students were expected to repeat experiments until they had reached the required minimum quality. This was only practical because of the additional lab hours for each experiment, including additional free hours where students could drop in if necessary to complete their experiments. Part of the grade for the lab reports was devoted to accuracy of the results of the experiment.

In addition to the reports, Chemistry 7L students were assessed by a lab final at the end of the class, which tested their knowledge of the procedures and the expected results of the labs. Although this was a written (multiple-choice) test rather than a lab practicum, it further pushed students to learn and retain the details of their experiments, even after the reports were turned in.

These different elements supported each other, encouraging students to feel like scientists doing actual science, while simultaneously teaching them the basic experimental tools and safety that would be required for their possible future research careers. While the workload of this class was significantly greater than that of the physics 1-Series labs (four units as opposed to two units), the 1-Series lab would not have been organized this way even if it were a four unit class. This

procedural and formal write-up focus was a much better fit for biology students in chemistry (where there is significant overlap in terms of procedures and practices) as opposed to those same students in physics.

Both the physics and chemistry lab classes were recently redesigned in an intentional way, with everyday practices supporting their high-level lab goals. Each class supported its own ideal lab frame for its own students. While course artifacts and surveys can help to assess student skills, they are a poor tool for examining individual students' own lab frames. How did students perceive these two lab environments? Did they view them as two basically similar lab classes with different lab report structures, or did they approach each class with a distinct lab frame? How close were these lab frames to the intended frame of the developers of these classes?

### **Students perceptions of chemistry and physics labs**

In order to begin answering these questions, we interviewed a small handful of biology students who had taken both Chemistry 7L and at least one quarter of the 1-Series lab sequence. The focus of the informal interviews was to see how the students perceived each lab class overall, and especially how the structures of the lab informed their lab frame.

For each of the interviewed students, the goal of chemistry lab class was clear: learn to perform in real life the procedures you had learned about in general chemistry, and present that information in a professional way via the lab reports. As Student A put it:

Basically any experiment or like any procedures that we've discussed in class before we had to demonstrate that during our actual lab class. So like precipitation,

titration, anything that's General Chemistry-related that you learned, you have to like learn how to do it.

Student B agreed, and expressed some frustration, since the focus on accuracy of procedure meant that experiments might have to be restarted if the quality level of the end chemical product was not high enough:

You work with a centrifuge and I was really surprised with how much of a pain it was to simply have to get your solution and have it spin, and then precipitate, and let that precipitate on the bottom and then have to extract the solution itself... So it was a lot of "you know it, but you need to be precise."

Still, that student found that the drilling on procedure had been quite valuable, especially since they found that directly relevant to the tasks they subsequently performed in their undergraduate biology research group.

Similarly, all of the students viewed the formal lab write-ups as training for a future career in biological or medical research. The value of this instruction was dependent on whether or not those students were preparing for a future in research, or being an actual medical doctor. Interestingly, Student A credited the chemistry lab class for their realization that they did not want to pursue research in biology:

So it is a class that has helped me, because a lot of people do apply to medical school and take the research path, and that class kinda helped me realize that I don't want to.

The chemistry lab class, then, successfully projected authenticity with respect to actual biological research for the interviewed students. While it may not be true that the chemistry lab class portrays the independent, self-directed scientific research accurately, the students currently engaged in undergraduate research felt that the lab was consistent with their own research. This is an impressive feat for an introductory lab class.

1-Series physics lab classes, of course, had none of that continuity with actual undergraduate physics research, either in content or in course design. Student A correctly assumes that this is because the class isn't trying to teach physics research practices to prospective physicists:

I feel like not many people taking physics 1-series usually go into physical research, so that's why they focus more on the content rather than... what physics research is like.

The students viewed the overall purpose of the physics lab class as a complementary way of teaching the new physics ideas from the lecture class. This allowed the results of experiments to be more surprising and memorable than in chemistry, since in chemistry the approximate results of the experiments were already known to the students because of previously taken courses.

In the "Shape Races" physics lab, students rolled different shapes (ball, ring, etc.) down a slope in several trials to determine whether mass, size, or shape influenced the results. Students performing the lab typically have not had instruction in rotational inertia yet, so the lab activity is exploratory. This experiment was the most vivid memory of multiple students, like Student A:

I think when we rotated all of those things in 1A, and then—I remembered I was shocked because we had to predict which one rolls faster, and I remember I had the wrong one.

Coincidentally, a very similar set of questions were later on that student's MCAT, and they credit the lab experience for making that portion of the exam easy. This is an optimal result of concepts-focused lab instruction for both instructors and students.

Student B described the same lab this way:

Watching the disks and hoops roll down – it was really interesting because I'd think "oh, the hoop would go faster" or something, and a lot of things really came to a surprise since a lot of them tied. So it was really nice to see what actually affected [race outcome] in the inertia equation

Here the student presents an internal loop of scientific discovery: hypothesis, experiment, unexpected result, and connecting that to existing principles. Because of students' relative lack of experience in the concepts, these moments are more likely in physics lab than in chemistry lab.

By contrast, students felt that the methodical, careful nature of some chemistry experiments, like titration, could be its own reward:

There was actually a lot of very strong acids that we had to use, a lot of colors, so I think that was fun. And then it was the most tedious lab, it was one of the most tedious labs, which I think was fun and challenging. So just because you had to titrate at a certain point, and if you didn't you had to start over. So you had to be the most patient, so that was the most – you had to be really patient, so it was kinda fun

If a goal of the chemistry class is to teach students how to perform procedures, the students need to spend time repeatedly practicing those procedures. While some tedium in such tasks is probably unavoidable, the selection of colorful chemicals used encouraged the student to continue carefully performing the experiment, to get that necessary practice. In most introductory physics labs (even for majors), learning the procedures is less relevant, which makes repeated trials seem more pointless.

This is another side of science: the methodical and careful recording of data and the preparing of experimental samples. Even if the ultimate result isn't a surprise to the chemistry students, the focus on a specific answer allows them to refine their experimental procedures and skills. The physics analogue to the chemistry lab class is not the physics lab class, but the machine shop, where the focus is on learning how to use the equipment with precision. In both cases, the ideal result is a trained student who is able to use the equipment safely to accurately make objects or concoctions that will help their research.

Even though both classes required students to determine experimental uncertainties, the focus and techniques were quite different. Student B describes the way that uncertainties were determined in chemistry:

For any measuring tool there were specific rounding you had to do. You had to go to the thousandths place [for volumetric flasks]. Make an assumption at the thousandths place and that would be your uncertainty there. And that's how it would work.

To them, these rules (along with rules for significant figures) were simple, repetitive, and would always work. By contrast, physics led to more confusion because of the necessity of propagating errors for unmeasured quantities. Because of the multi-step nature of the process, there were more failure points where mistakes could be made and points lost.

This additional complication also extended to how the students interpreted the results of their experiments. Since a significant goal of the chemistry class was to train procedures, the sense-making part of interpreting data was basic, as Student B describes:

The procedures would outline what you were looking for and so to come to your conclusion you would have to know specific equations and calculations and concepts and apply them all together into your data. And then you'd get a specific value and you'd relate that to some charts you were given and see "oh, how much pH does this have? Oh obviously this is my solution".

While mistakes could be made (if the color of the final solution was wrong, for instance), the evaluation of the final result was simple: does it fit the described characteristics or not? While this is an excellent fit for a class teaching experimental procedure, this is not scientific sense-making or data evaluation. Nevertheless, in experimental science data does need to be collected with a certain precision, and that is a legitimate skill that should be trained as a part of an undergraduate degree.

For her physics lab class, by contrast, Student B draws a direct line to the scientific method:

A lot of it was if you wanted to see if, how momentum acted in a specific way you'd do different trials, you'd change a variable, you'd have something controlled.... A lot of it's just scientific procedure: you'd find your errors and you'd see what works and what doesn't work and you'd figure out the concepts. Just like in the 1BL lab I did the other day with just working with charges, charging, rubbing up on PVC pipe and seeing how much it would attract the aluminum foil.

In most cases, the physics lab data comparisons weren't accuracy comparisons at all, since there was no expected result with which to compare one's data. After all, no one actually knows how much a particular charged strip of tape should attract a given PVC pipe without actually making the measurement! This resulted in an environment where students were expected to work out some of the main ideas themselves, rather than always being expected to know what they were in advance.

The structural differences between chemistry and physics lecture classes matter, here: there was less value in exploratory labs for chemistry, since the students (in principle) were already familiar with the basic concepts in their previously completed general chemistry class. And while exploratory lab class are well-regarded with good reason in the PER community, they cannot easily coincide with the focus on safety required for a chemistry lab.

Overall, the interviewed students were able to perceive the design goals present in each class. For them, the weekly class and assessments complemented what they saw as the overall focus of each class: the methodical, procedural chemistry labs culminated in a predictable, comprehensive lab report that emulated research norms in their science; by contrast, the more discovery and concepts-focused physics labs were assessed in a way that was centered on the ideas they were learning.

Nowhere was this clearer than in Student B's description of their own ideal mental states in chemistry and physics classes, respectively:



I'm on the clock. Time is my enemy, and I need to make sure that I'm as precise, clean, and well-read on what I'm going to do today. That's my mindset. I try to make sure I don't make mistakes, I don't break anything. If mistakes happen, I correct them as soon as I can. That's what it is in chemistry lab: I'm trying to be as efficient with my time as possible.

I need to be taking notes for my conclusion. It's not just saying "my notes are in the questions I've answered". No, it's how well did I understand the concept: can I explain it to myself again without the TA there? Where am I going with this? Does this make sense? Do I understand conductors and insulators? Does this experiment make sense at all? Why is it like this? I'm asking a lot of "why" questions, how, and where I'm going to go with it. I think that's what physics lab is. It's not so much an emphasis on how efficient you are, it's how well-read and how well you can conceptualize something.

Both of these lab frames are appropriate in the correct context for a researcher: if I'm being too epistemological and metacognitive while I'm measuring a solution or machining a part, I could injure myself! Rather, the approaches of the two classes present different but complementary perspectives on how science is done. It is a minor miracle that both classes did complement each other so well without any intentional planning!

### **Physics service: a biologist's perspective**

Although it is beneficial overall that the physics and chemistry lab classes were overhauled in such a comprehensive way, they are ultimately service courses for not only the biology students but the biology department as a whole. Despite feeling that my project was valuable, I remain a bit bemused that it was possible for me undergraduate lab material for biologists without consultation with their own department! In an effort to address this, I spoke to Dr Gabriele Weinhausen, a biology professor at UCSD currently working to improve undergraduate biology education, including the creation of a new lower division biology lab that would be taken early in

the educational sequence. This class could help to address the central awkwardness described above: that the chemistry and physics lab classes are filling a void of hands-on scientific thought in lower division that has been left by the biology department for their own students.

Traditionally, she describes [43], biologists at UCSD viewed math, chemistry and physics as the foundation for biology instruction; consequently, there had actually never been a lower division biology lab in UCSD's history. Lower division biology has traditionally a focus on facts and memorization that in many ways is somewhat frustrating for a modern, student-centered approach towards learning. Without a lower division lab class, she viewed physics as providing some of the raw tools and practice to allow students to appreciate conceptual understanding and assessing whether the outcome of an experiment makes sense. Unlike material in a biology textbook, physics lab exercises cannot be memorized; moreover, they show a connection between the tools of an experiment and what the experiment was trying to measure that is hard to encapsulate in a purely written text.

The laboratory goals and execution of my overhaul of the curriculum do line up with her perspectives on the value of physics lab education. Still, I've come to appreciate as a part of this project that my approach towards physics lab instruction may only be passingly helpful to establishing the appropriate frame for actual biological lab instruction. While the basic principles might be appropriate (relying as they do on educational principles that transcend any particular science), physicists and chemists cannot replicate the value of biology students learning to emulate actual biologists' scientific process and thinking. Indeed, Dr Weinhausen created the lower division laboratory class for this very reason. As she described her ideas for this lab class (which included a research connection to several faculty labs, a focus on mirroring real research papers and even an undergraduate research journal that students could publish in) I was reminded of

Kung's scientific community labs in physics. Here also was the focus on emulating the broader community of science researchers that students may eventually join.

### **Future work and lessons learned**

In places like UCSD, where undergraduate biology lab courses follow chemistry and physics lab courses, it is crucial for those non-major lab classes to complement the eventual biology instruction, both in conceptual content and in terms of general science skills. While the chemistry and physics lab classes managed this (more or less by coincidence) at UCSD, a more intentional approach is more likely to produce useful results. My specific suggestions from my experience follow:

#### **1. Center biological content, with consultation with the biology department.**

While preparation for the MCAT can be a practical guide to the development of physics concepts used, this barely serves the biologists that are not planning to become doctors. Even if lecture content varies based on the instructor, the lab experiments and questions can be targeted towards topic areas that the biology faculty know will be useful.

For example, study of fluids at UCSD is sometimes included at the very end of mechanics in 1A, but is often cut for time limitations. There is currently no fluid based lab in the course. But fluid flow (as studied in for example, Bernoulli's equation) is an important topic area to study for biological applications such as how blood flow speed relates to capillary size.

Student A's 1A lecture class did manage to cover fluid flow, and they found it very useful in explaining the reasons for the rules they later learned in biology class:

It helped me the most with the circulatory system, just because we had to learn understanding why blood flows in constricted areas, like how that rate changes, and then they presented the formula. Having an idea of what that formula is, being exposed to it so much before, kinda made understanding it easier, rather than [just] accepting the fact that blood flows faster.

This is an excellent outcome: with physics filling in the “why” to back up an important fact in the biology class. By intentionally seeding classes with the physical phenomena that are most important for biologists, these experiences can be made more common. The best way to find these areas is via direct consultation with the instructors teaching introductory biology. Implementing these concepts within the lab class is a way of ensuring that students definitely experience these key topics, without relying on instructors reaching the end of the course material. Of course, this is best achieved if the labs themselves are exploratory in cases where the students may not have covered the concepts in class beforehand.

While I'm not suggesting that physics course content should be entirely dictated by the biology classes for 1-Series classes, applications especially can be developed with biology in mind. In some areas, this overlap is natural: for example, the use of lenses in optics is easily tied to prescription lenses, which is helpful for future opticians. For a time we were considering cutting some pre-lab lens questions about the diopters system (because it was extraneous from a physics-only perspective), but chose to keep those questions on as a relevant biological application. This also increases the relevance of these classes to the students, helping them understand why they are taking the classes at all: all of our interviewed students felt more connected with the material in 1BL and 1CL (E&M, waves, optics, etc.), because it was much more relevant to their actual areas of interest. At some schools, lab procedures for non-majors may have been adapted from lab

procedures for the major or engineering sequence with the more difficult math removed. Some staple lab experiments (e.g., momentum conservation using carts on air tracks) that might be less useful conceptually for biology students to retain might be substituted in favor of fluid labs to better reinforce the key principles that biology instructors would like their students to be more conversant with.

## **2. Discuss scientific method / lab practices goals with chemistry / biology**

Physics lab treatment of error and uncertainty can potentially undercut subsequent treatments of the same in other lab classes. Although I clearly think that a focus on estimation, propagation of uncertainties, and the use of data ranges to draw conclusions has clear and demonstrable value, the emphasis should change depending on the specific needs of the later biology lab priorities.

For example, our treatment of uncertainty in 1-Series did not focus on differentiating between accuracy and precision, and consequently elided the difference between systematic and statistical errors. While this was a defensible choice in viewing those labs in isolation, we would have prioritized differently if that distinction was crucial in later biology lab classes.

Although I eventually reached out to lab designers in both chemistry and biology to present the changes we had made in modifying our lab courses, ideally that discussion would have happened before the implementation of this project altogether. Indeed, broad changes of the type that were made probably shouldn't be possible in the absence of at least informal consultation of these other departments. After all, 1-Series labs are service courses! Despite the presumptive expertise of physics lab course instructors when it comes to physics concepts and lab structure,

they are certainly not experts in what practices are most useful for biology students to learn, in the context of their lab education in biology.

While this sort of coordination is not easy (it has proved difficult enough even for lecture and lab instructors in the same department!), it is necessary to help ensure that these lab service courses remain relevant to both the students and the departments that they serve.

### **3. Potential lab course pedagogy changes**

One major implementation detail served to limit the conceptual scope of what was changed in the uncertainty content of the labs. Because 1-Series lab classes were major requirements but not prerequisites for any of the other major classes, students often postponed some of the lab classes until late in their degree. This meant that a significant fraction of the students (~20-25%) taking 1BL, for example, would not have taken 1AL in the previous quarter, even for the “on-sequence” lab courses. If the uncertainty content of 1BL relied on students having taken the new pedagogy for 1AL also, that group of students would have been left in the lurch. Consequently the decision was made to treat each course separately: each lab class would begin with measuring, estimation, and determining a spread, and only later move on to propagation. This also allowed changes to be phased in gradually without rolling out changes to three classes at once.

With the basic structure of content, training, and skill building in place, treating the entire 1-Series as a single course for uncertainty would allow more room for sense-making about uncertainty in the labs, rather than a focus on calculating spreads and using them to propagate uncertainties. What follows are some untested ideas that I would like to try in that context.

Despite the special attention placed on edge cases in the material provided to students (see Figure 2-5 and Appendix i), after instruction only a small percentage of students answered survey questions about ambiguous data with the correct “not sure” answer. For students to gain more practice interpreting these situations, they would need to record ambiguous data in the lab, and would likely need some in-class discussion and work time to help. Of course, there is no guarantee that students’ lab data will be an edge case at any point in the lab; moreover, asking for their TA’s help in those cases might not be helpful anyway, considering some of the novice uncertainty ideas that some teaching assistants still retain (e.g., in estimating uncertainties).

Taking some time out of one lab to analyze a data agreement problem (with pre-created data which is an edge case) could help to mitigate that. With simulated data (as in Figure 4-10), students would always be sure to encounter a problem where the edges of the data just barely touched, and could have the time to work the answer in a group context. That sample problem’s analysis would be a key topic of that week’s TA training, to help ensure the appropriate guidance is provided. Similar problems could be placed early in the quarter of 1BL and 1CL to remind returning students (and help to familiarize new and transfer students) of the proper lab frame to take in 1-Series when comparing ranges.

Another regrettably small stream of student responses to encourage are those that saw uncertainty as having a significant role in science practice, especially in the communication of scientific results. One option would be to spend a lab working on context-heavy problems where groups serve as the decision-maker in how to communicate uncertainty to other scientists or the general public. Each group would be given a set of data either in graphical or tabular form along with a short paragraph explaining the context. Some of the situations would be low-stakes (how long to oven-cook a pizza to get the crust just crispy enough?) whereas others would be higher

stakes (what should the safety standards be for asbestos poisoning?). While for the low stakes problems it would be fine to treat the upper and lower limits of the range as is (since slightly over and undercooking a pizza is not so important), a larger safety margin would be appropriate for the higher stakes problems where literal safety is at stake. The activity would be structured so lab groups would get 45 minutes to discuss the situation and come to their conclusion; 45 more minutes would be spent with the groups presenting their results and why; with the leftover minutes for general TA closing comments. This sort of lab activity would be properly performed during the last lab week in one of the quarters, since there is less need for a post-lab assessment. In addition to giving students the chance to see how uncertainty manifests in the real-world, this sort of problem could highlight conceptual connections between biology and physics based on the choice of situation.

In the survey questions on simulated computer data acquisition shown in Figure 4-5, some students spontaneously reported the value of the quantity using a range based on the upper and lower limits of the graph. In the arrow measurement task, students almost never reported a spread to describe the length of the arrow, but only did so when an uncertainty was specifically requested. One possible reason for this difference is because the graph contains many data points that are shown all at once, allowing even a casual observer to see that the quantity varies; by contrast, a tool like a ruler measures only one number at a time, making the idea of a spread less immediately accessible. There may also be a difference between graphical display of uncertainty and lists of individual numbers, as seen in Kung's range comparison questions in Figure 4-8. Finally, there may be a distinction between graphical results and the standard average  $\pm$  uncertainty format that is a standard in communicating results, as seen even for experts as seen in Figure 2-3. How much do the different representations of uncertainty matter in communicating results?



A systematic study of how effectively these representations communicate ideas of uncertainty would be helpful. The goal would be to determine which representations are better at priming the idea of spread, especially for students new to uncertainty. This could serve to change which tools are used when in courses like the 1-Series, where the choice of representations has mostly been based on the convenience of the instructor of the lab course rather than any particular pedagogical reason. Cues can be taken from Epp and Bull's review [44] of graphical representations of uncertainty and how effectively they communicate variability to the observer. Lab tasks could be designed to work with different representations specifically, rather than reducing all data inputs into the uncertainty  $\pm$  framework.

Finally, cooperative labs are an inherently social space. Data, then, is something of a group responsibility, and I am interested in how measurement uncertainty factors into this, especially in environments like the 1-Series where uncertainty values are (intentionally) high. I have observed students discussing whether or not to drop (good!) data points because they are perceived as too far away from previously measured data points or the expected result. How often do these discussions happen? Are they more likely to happen with low precision or low accuracy experiments? How do groups resolve differences of opinion about estimates of uncertainty? One way of learning more would be recordings of in-lab conversations about uncertainty, perhaps taken in the style of Kung and Linder's study of in-lab metacognition [37]. In that case, student conversations was classified on whether it was off-task, logistical, or sense-making. Although their goal was to see how often metacognitive questions transitioned students into a sense-making frame, the same basic technique could be applied to measure group conversations about data, especially how often students record or discuss uncertainty. This could be used as a way of assessing how effective the different representations are of encouraging discussion or reasoning

about uncertainty in the actual lab context. This would allow the group context of actual lab measurements of uncertainty to be recorded, rather than just the approximation based on students working alone on survey items that have been the main assessment tool of this study.

#### **4. Comments on the use of physics attitudes**

Attitudes towards physics (or more generally, science) as measured by tools like the CLASS are used as another measure of expertise. The brains of experts in a given field are different than those of novices; not only do they contain more factual knowledge, but also they organize that information hierarchically, they monitor their own understanding metacognitively, and so on [45]. It makes some sense, then, to measure how and if the opinions, feelings, and self-reported practices of experts differ than novices. Indeed, the CLASS found in its survey of faculty members a certain consonance of approach: experts think and feel about their field in a certain way, largely different from their students.

But attitudes surveys are different from more operations-based measures of expertise, where the novice is unable to reproduce the same actions as the expert. A new student cannot solve a complex problem that their teacher finds trivial in the same way that a beginning piano student cannot play any piece well; they have not yet learned the technique and put in the practice to do so. Attitudes surveys are different: Adams et al. found in the initial CLASS study that novices can, in fact, answer the study like experts if they want to [24]. Although this was presented in the context of validating the instrument (we know students aren't just answering like their professors want because there is a difference between what the students say for themselves vs what they say

their professor thinks), it makes the interpretation of the instrument more complicated. What does it mean to be an “attitudes expert” according to a tool like the CLASS?

Even though physics content experts answer the CLASS in a certain way, they aren’t the only ones that are attitudes expertw. Introductory classes also have attitudes experts, even in non-major courses like the 1-Series; despite their expert attitudes, of course, they do not have the content knowledge (or presumably, expert brain organization and approach) of their expert professors. Are they *potential* professors of physics? Are they future biology professors and the attitudes towards science are transferrable? Are they answering to please their professors, or as they feel they should?

One feels that the original intent of these studies was to make these questions irrelevant through instruction. Although most of the students might begin class as attitudes novices, after a course in science they would presumably be more expert-like on both a content and attitudes basis. But for the CLASS, the opposite result was typical: students generally became less expert-like in their attitudes as a result of taking the class. Where do physics attitudes experts come from? Do they develop during the degree in ways that aren’t attributable to classes? Or are the eventual physics majors just the students that were attitudes experts in their introductory classes?

This last idea was studied by Perkins [46] at the inception of the CLASS. Indeed, she found in her longitudinal study of intended and eventual physics majors that CLASS scores were higher for those that ended up majoring in physics, even in their introductory classes. This was true even for students that did not originally intend to be physics majors; they still entered college physics with “very expert-like beliefs.” And although Gire’s study [28] of CLASS results at various levels of undergraduate and graduate work wasn’t quite longitudinal she found something suggestive: as

the upper-division populations became smaller and smaller as students dropped out, their average attitudes score became higher and higher. Indeed, the very expert-like attitudes of the graduate students (88.1% favorable) were from a population of only seven.

Besides retention, the graduate student results may be higher than those of fourth-year majors for a different reason. At this point, it makes sense for students to actually be self-identifying as professor-like physics experts. The phrasing the original CLASS study [24] used to differentiate personal opinions from the (hypothesized) professor's opinions was: "What would a physicist say?"; and "What do YOU think?". When graduate students answer the CLASS, they answer just as their physicists would, because at this point "YOU" *is* a physicist. Hazari, et al., developed a framework [47] defining self-assessed physics identity based on four factors: "(i) interest (personal desire to learn/understand more physics and voluntary activities in this area), (ii) competence (belief in ability to understand physics content), (iii) performance (belief in ability to perform required physics tasks) and (iv) recognition (being recognized by others as a physics person.)" Just being accepted into graduate school would certainly strongly trigger each of these factors.

What about the students that take the early major classes but are not found in the year 4 classes? Are they getting weeded out by the difficulty of the material, causing them to doubt their competency? Were they just not that interested in physics? There is a potential danger of viewing these results as proscriptive, not restrictive; they do not view themselves as physicists because they are not the people that belong in physics.

Why might someone who can do physics still feel that physics is an unwelcoming place for them? Some of this is simple faculty demographics: if you are an underrepresented student,

you are unlikely to be taught by faculty like you. As an undergraduate in physics, I was only taught physics by white and Asian men, for example. My experiences are not atypical: as of 2010, only 26% of new faculty hires were women, with 47% bachelors-granting departments having zero women on the faculty at all [48]; in 2012, only 2% of physics faculty were African-American, and only 3% were Hispanic [49]. The intersection of race and gender is even more dire: at PhD granting institutions in 2012, there were only 14 African-American women faculty, and only 19 Hispanic women faculty. These represent 0.2% and 0.3% of total physics faculty at those institutions. The American Institute of Physics webpage from which these statistics come does not include any mention of Native American faculty members, nor breakdowns based on LGBTQ status or disability status. Even as a relatively privileged person (male, cis, non-disabled, straight, but non-white and non-Asian), my odds of being taught by someone like me at my undergraduate school were extremely low.

Of course, there is more to it than not feeling represented by the faculty who teach you. Underrepresented students are more likely to be exposed to microaggressions (“brief, sometimes subtle, everyday exchanges that either consciously or unconsciously disparage others based on their personal characteristics or perceived group membership” [50]) because of their marginalized status(es). Although microaggressions have only recently begun to be studied in the context of science, it is clear how they can contribute to a feeling of not belonging in science; and the relative homogeneity of physics faculty make the chance of hearing them from an authority source more likely. Hazari’s study [47] of high school students found that self-identification as a physicist was highly correlated ( $r=0.7$ ,  $p<0.001$ ) with whether their teachers viewed them as a “physics person.” Consequently, microaggressions from teachers may serve to undermine the self-identification as a physicist. On the other hand microaffirmations work in the opposite direction: female students had

significantly higher physics identity if their high school class had a discussion of women's underrepresentation in physics. Their self-identity affects their goals, since students that viewed themselves as physicists were much more likely to plan for a career in physics. While these correlations were only measured in high school students, it is hard to imagine that similar factors are not at work for undergraduate students.

What does this have to do with the CLASS? The survey is designed to measure not only abstract ideas about physics ("knowledge in physics consists of many disconnected topics") but how the answerer feels about physics ("I think about the physics I experience in my daily life"; "I enjoy solving physics problems"). The authors have resisted attempts to remove these items [51], and other items that do not have an expert consensus. These Personal Interest items were found to be a strong predictor of overall interest in physics [29] and were also answered by women in a less expert-like way[52]. While the CLASS does not directly measure microaggressions, harassment, etc., as a survey measuring affect it may still quantify some of their effects: decreased enthusiasm for the field. A future study exploring the connections between physics identity, microaggressions, and enthusiasm may be anchored in part by the CLASS, which is the most common way of measuring physics affect by far. If a correlation can be found between the general attitudes results and changes in physics identity, it may be possible to infer those effects from the broader pool of CLASS results.

## Appendix I: Uncertainty and Error Document (Winter 2013 1AL)

### Introduction

The values that you measure in the lab are not perfectly accurate or precise. Each measurement you make has an uncertainty associated with it that depends on several factors, including the precision of the measuring instrument, the variability of the measured object and other complicating factors. You will need to report an uncertainty for each set of measurements you make in the lab; how to come up with that uncertainty, what it means, and what to do with the uncertainty is the subject of the rest of this document.

For this class, we will use a slightly informal definition of uncertainty. The **uncertainty we use in this class defines the likely range of a set of measurements taken with the same conditions**. When you quote a measurement and its uncertainty as  $210 \pm 8$  m/s, that statement implies that if you measured the speed of an object under identical conditions many times, almost every measurement of its speed would be between 202 m/s and 218 m/s. (For those of you that have some kind of statistical training, the uncertainties in this class will almost always be around 2-sigma or 3-sigma, rather than the standard deviation)

**Use this Uncertainty and Error document as a reference guide**, as each lab will present different challenges related to measurement and uncertainty.

### An Uncertainty for Each Measurement

One way to figure out the uncertainty for a given measurement is to **make the measurement multiple times, and determine the spread of the data**. For example, if you measured the length of a table three times and got 125 mm, 128 mm and 121 mm, the average value of the length would be 124.7 mm. The spread of your measurements would be from 121 mm to 128 mm. Take the difference between the average and the furthest measurement away from the average; that difference is an estimate of the error. In this case, that error would be  $124.7 \text{ mm} - 121 \text{ mm} = 3.7 \text{ mm} \rightarrow 4 \text{ mm}$ . You would quote this measurement as  $125 \text{ mm} \pm 4 \text{ mm}$ .

**Sometimes, you will be able to figure out the measurement uncertainty of a device by playing with it.** For instance, suppose you are weighing an object with a balance. When the two weights are almost balanced, a small change in the mass applied can make the beam teeter significantly. By moving the smallest increment back and forth and seeing how the beam moves, you can estimate how much the weight can change and still leave the scale basically balanced. You should be trying to answer the question “how much could I move this sliding weight and still believe the two sides of the scale were balanced?” That amount is the measurement uncertainty of the device. This technique will be most useful when you have fine control of some part of the measuring device, and when it is somewhat complicated.

Sometimes, you will be using an **electronic data acquisition system**, which will typically make the same measurement repeatedly, often many times per second in the case of LoggerPro. In this case, **as long as conditions aren’t changing, the range that the data spans over a few seconds of measuring is the same as the measurement uncertainty of the device for that measurement.** Usually, just by looking at the visual display, you will be able to figure out the range of the measurements, and then work out the uncertainty as if you had measured the spread itself (ie, by dividing it by two).

When multiple measurements aren't practical, you can make a **reasonable estimate of the uncertainty.** This estimate should represent how confident, quantitatively, you are in the result. Choose your estimate of the error so that you create a spread that you believe will contain the actual value of the measurement. Don't just choose the smallest significant figure from the measuring device, since that doesn't tell you how accurate your measurement is! For example, let's say you measured the length of the lab wall with a small, 10 cm ruler. Even though the ruler can measure distances to 1 mm, the measurement of the wall won't have that accuracy: you would have to lay down the ruler parallel perfectly many times in a row at exactly the correct place. Your error estimate should be an estimate of all of these imprecisions and inaccuracies. **Remember, the uncertainty implies something about the largest possible and smallest possible values a quantity could be.** In the example above, the statement that the table is  $125 \text{ mm} \pm 4 \text{ mm}$  is interpreted as “the table's length is almost certainly between 121 mm and 129 mm; it could be anywhere within that interval.

In any case, the error should always have the same units as the quantity that is being measured. **Every measurement you make in the lab must have an estimated or calculated uncertainty associated with it.** Sometimes, you will not have directly measured a quantity in the lab (ie, you calculated it from an equation), but you still need an uncertainty for it. See “Propagating Uncertainties” below.



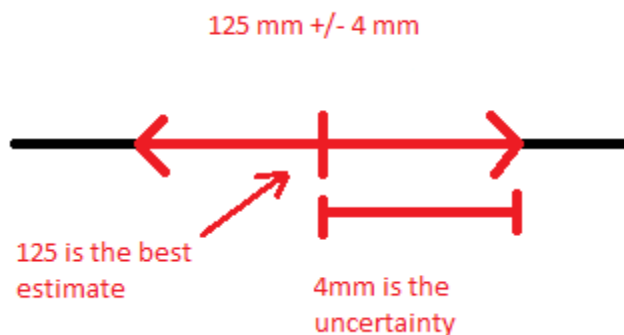
## Comparing Data

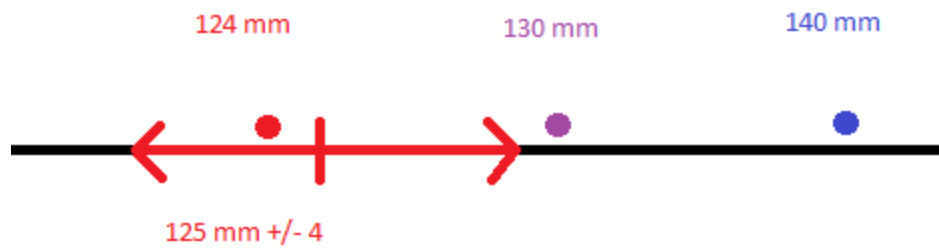
Sometimes, you will be asked to compare data to a theoretical prediction. **The data is consistent with the theory if the theoretical prediction is within the spread of the data.**

If the theoretical prediction is very close to the edge of the range, then you can't be sure if they are consistent. For example, say that the theoretical value for the length of the table was 140 mm. This is far outside the spread (121 mm – 128 mm), so you would be confident that the theory is inconsistent with your data. If the length of the table were 124 mm, you would be similarly confident that the theory and data were consistent. If the length of the table were 129 mm or 130 mm, however, you wouldn't be particularly confident either way. See the picture below for an example of how to judge the consistency of a range with a prediction.

In a similar way, one set of data can be compared to another set of data. In this case, compare both ranges. If they strongly overlap, the agreement is good. If the ranges don't overlap at all, and it's not close, then the two sets of data are inconsistent. If they just barely overlap or just barely don't overlap, you can't be sure, and you should say so in your conclusion.

Sometimes, you will compare your measurements to known constants, like the gravitational acceleration  $g$ . It is okay if you find your data to be inconsistent with these known values! **Always report what your data tells you, NOT what you think the answer should be.** Your labs will be graded accordingly.





124 is consistent with the measurement

130 is borderline; can't say much either way

140 is inconsistent with the measurement

## Propagating Uncertainties

Often, **you will measure one variable (and its uncertainty) in an equation and want to know the uncertainty in another variable, which you did not measure.** For example, you might have measured how long ( $\Delta t$ ) it takes a ball to roll a certain distance ( $\Delta d$ ) at a constant velocity ( $v$ ), and you might want to calculate the uncertainty in that velocity. You should use the following technique in this course to calculate such errors. This technique is not statistically precise, and results in an overestimate of the uncertainty for that quantity.

**The goal of this method is to determine the possible range of the unmeasured quantity, and then use that to get its uncertainty.**

1. Solve the equation for the variable whose uncertainty you want to know.

In this case, the equation comes from kinematics:  $\Delta d = v * \Delta t$ .

Isolating the velocity, we have  $v = \Delta d / \Delta t$

2. Calculate the variable whose uncertainty you want to know, ignoring the uncertainty for now. This is your **best estimate** for the variable.

Let's say  $\Delta t$  is measured to be  $0.42 \text{ s} \pm 0.02 \text{ s}$ , and  $\Delta d = 1.80 \text{ m} \pm 0.13 \text{ m}$ .

$$v = \Delta d / \Delta t = 1.80 \text{ m} / 0.42 \text{ s} = 4.2857 \text{ m/s}$$

3. Calculate the maximum value of the variable that is consistent with your uncertainties. To do this, calculate the variable again, but this time, replace each measured variable with itself **plus or minus its uncertainty**,  $\delta$ . You will have to determine the sign separately for each of the measured variables; apply the sign that causes the calculate variable to increase.

We have  $v = \Delta d / \Delta t$

We want to figure out the maximum possible value of  $v$ , which we call  $v_+$

If you replace  $\Delta d$  in that equation by  $\Delta d + \delta d$ ,  $v$  would increase. Thus it is the correct choice for calculating  $v_+$ .

If you replace  $\Delta t$  in the equation by  $\Delta t + \delta t$ , note that  $v$  would actually decrease. We don't want that, so we should use  $\Delta t - \delta t$

Plugging in for our example:

$$v_+ = (\Delta d + \delta d) / (\Delta t - \delta t)$$

$$v_+ = (1.80 \text{ m} + 0.13 \text{ m}) / (0.42 \text{ s} - 0.02 \text{ s}) = 1.93 \text{ m} / 0.40 \text{ s} = 4.8 \text{ m/s}$$

Note that, as expected, this is higher than the best estimate of  $v$ , which was 4.825 m/s. This is our **overestimate**.

4. Repeat step three, except now trying to calculate the minimum value of the variable. (The signs should all switch, but you should double check.)

If you replace  $\Delta d$  in that equation by  $\Delta d - \delta d$ ,  $v$  would decrease. It is therefore the correct choice for calculating  $v_-$ .

If you replace  $\Delta t$  in the equation by  $\Delta t + \delta t$ ,  $v$  increases.

So:

$$v_{-} = (\Delta d - \delta d) / (\Delta t + \delta t)$$

$$v_{-} = (1.80 \text{ m} - 0.13 \text{ m}) / (0.42 \text{ s} + 0.02 \text{ s}) = 1.67 \text{ m} / 0.44 \text{ s} = 3.7955 \text{ m/s}$$

Note that this is less than the average value of  $v$ . This is our **underestimate**.

5. Take the difference between the overestimate and underestimate, and divide by two. This is the **uncertainty** for the calculated variable. Make sure to express your answer with the right number of significant figures.

$$(v_{+}) - (v_{-}) = (4.825 \text{ m/s}) - (3.7955 \text{ m/s}) = 1.0295 \text{ m/s}.$$

$$1.0295 / 2 \text{ m/s} = 0.51475 \text{ m/s} = \delta v$$

6. Make sure each of your calculated values uses the correct number of significant digits. Generally, the uncertainty should be only reported to one significant digit, unless that digit is a one, in which case the next digit should be included (ie,  $\delta v = 0.12732$  would be rounded to 0.13, with both digits kept).

$$v = 4.2857 \text{ m/s} = 4.3 \text{ m/s}$$

$$\delta v = 0.51475 \text{ m/s} = 0.5 \text{ m/s}$$

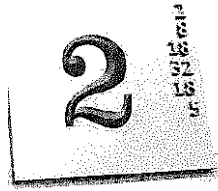
7. Report the calculated uncertainty along with the best estimate it was calculated for.

$$v = 4.3 \text{ m/s} \pm 0.5 \text{ m/s}$$

You should always quote data in the form: best estimate  $\pm$  uncertainty.

The above technique for calculating uncertainty is approximate, not perfect; it will tend to overestimate the uncertainty. If you know more precise methods of calculating uncertainties, you may use them, but be sure to make this clear in your conclusion. If you're interested in learning more rigorous methods at calculating uncertainties, [An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements](#) by John R Taylor is highly recommended.

## Appendix II: Waves Lab



# INTRODUCTION TO WAVES

---

## INTRODUCTION

Waves appear in many different places in real life and in 1CL: springs, strings, water, sound and light. Despite the different media involved, waves are described by the same basic properties: wavelength, frequency, and velocity. You will explore waves of several different types in this and subsequent labs.

## REQUIRED READING

- Chapters 13.1, 13.2, 13.3, 13.4, 13.6

Pay special attention to the distinction between the transverse velocity of a point on a wave with the speed of the wave itself, and what determines the speed of a wave, and what determines the velocity of wave. Also look at figure 13.19, and make sure you can tell the difference between longitudinal wave and transverse wave patterns.

Also read the learning goals below. Once you've finished all the work for this lab, you should be able to:

- make longitudinal and transverse waves using different materials, and be able to tell the two types of waves apart
- explain what sets the velocity, frequency and wavelength of a wave
- physically modify a wave you've made to change some of its properties, and explain which of its properties you can't change

## **B. MAKING LONGITUDINAL WAVES ON A SLINKY**

- B1. Stretch the slinky along the floor as in the previous section of the lab. Keeping one end fixed, move the other end about 5 cm (2 inches) in and then back out, in the direction that the slinky is stretched. The motion shortens the slinky for a very short period of time and sends a pulse down the slinky. Can you identify where the slinky is compressed?
- B2. Watch the piece of colored tape as the pulses go by. What direction is the tape moving? Is it different from the waves on the slinky in part A? Which way is the wave moving? How are these "longitudinal waves" different from the transverse waves in part A?
- B3. What happens after the pulses reach the stationary end of the slinky? Can you tell if the waves reflect off the end, or does something else happen?
- B4. Now have both ends send in pulses. What happens do the two pulses when they meet? Is it the same thing that happens when two transverse waves meet?

## **C. WATER WAVES**

In this part of the lab, you'll be creating waves in water and measuring their frequency, velocity and wavelength.

- C1. Locate the tub filled with water. Each member of your group should try generating waves in the water by using the ping pong ball. The ball floats in water, so by pressing it down to the bottom of the tub, you can create waves in the water. Practice making a consistent rhythm of waves.
- C2. Assign one member of your group to be the wavemaker. Hold the ball in the middle of the tub and try to produce about a wave a second, with a consistent pace. Try producing sets of ten waves, counted out by the person making the waves. The goal of the wavemaker is to get a consistent time with each set of ten. Time a few sets of ten, calculating an average duration for each set of ten, and calculate the uncertainty between the sets. When the wavemaker feels comfortable with that rhythm, that person should keep producing sets of ten while the other group members observe the waves in subsequent sections. Keep the same rhythm until otherwise instructed.
- C3. What is the frequency of the waves being produced? Calculate the frequency of the waves from the data in C2, including its uncertainty.
- C4. Find a good angle where the individual wave crests can be seen. How do the waves move after they are created? Do the crests ever catch up to each other? Are the crests moving at the same speed? What happens when the waves hit the sides of the tub?
- C5. Now you will measure the speed of the waves. The wavemaker should produce waves in sets of ten, counting out each time the ping pong ball hits the bottom of the tub. One group member should be responsible for timing the waves, and should take an angle where the waves crests are easily seen. Start the stopwatch when the tenth wave is produced, and stop it when that wave reaches the edge of the tub. (It is easier to follow the last wave with the eye than to track all of them at once.) Try this a few times, until you

## D. WAVE MACHINE

The wave machine is a look inside the inner workings of any wave medium. The “particles” of the medium are the balls at the ends of the sticks.

- D1. Move the particle on one end up the wave machine up and down. What happens to the other particles on the machine? Why does moving one affect the others? What kind of wave does the wave machine produce, transverse or longitudinal? How can you tell?
- D2. Make waves with a small amplitude by using the wave machine. How long does it take for the particle you are moving to get back to its starting location? How far does the particle have to travel to get back there? Does the particle move in the x-direction (ie, horizontally towards the next particle) or in the y-direction? How fast is the particle itself moving? Calculate this and label it as the “particle velocity”.
- D3. Make some more small amplitude waves, but this time watch the waves as they travel down the wave machine. How long does it take a wave to reach the end of the machine? Calculate this and label it as the “wave velocity”. Is it the same as the particle velocity?
- D4. Now make large amplitude waves and repeat your measurements of D2 and D3. Is the particle velocity of these large amplitude waves larger than for the small amplitude waves? What about the wave velocity? If you make a small amplitude wave and immediately thereafter make a larger amplitude wave, will the large amplitude wave catch up to the small amplitude wave? Try it and see.
- D5. The wave machine is a useful visual aid for explaining properties of waves. Imagine you wanted to show basic wave concepts to a friend of yours who hasn’t taken the class using the wave machine. Devise ways to address questions using the wave machine:
  - What is the period of a wave?
  - What is the difference between amplitude and wavelength?
  - How is the motion of a wave and a particle in the medium different?

Appendix II, in full, is a reprint of the material as it appears in Physics 1CL Lab Manual 5<sup>th</sup> Edition 2014. Schanning, Ian; Salamon, Joe; and others representing the UCSD Physics Department, bluedoor LLC, 2014. The dissertation author was the primary writer of this manual.

## Appendix III: Surveys

### Pilot Study (Summer 2012): Pre-Instruction

#### Multiple Choice

Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree

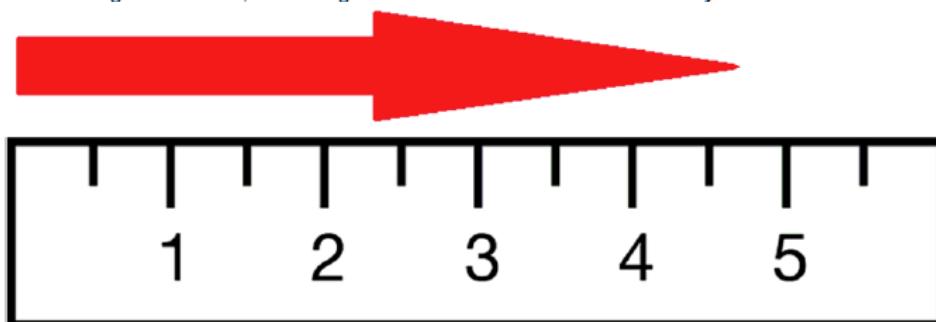
1. Problem solving in physics basically means matching problems with facts or equations and then substituting values to get a number.
2. In this course, I do not expect to understand equations in an intuitive sense; they just have to be taken as givens.
3. 1AL reinforced my understanding of physics concepts from lecture or the book.
4. Physical laws have little relation to what I experience in the real world.
5. In doing a physics problem, if my calculation gives a result that differs significantly from what I expect, I'd have to trust the calculation.
6. The most crucial thing in solving a physics problem is finding the right equation to use.
7. Physics is related to the real world and it sometimes helps to think about the connection, but it is rarely essential for what I have to do in this course.
8. The main skill I got out of Physics 1A was learning how to solve physics problems.
9. When I solve most exam or homework problems, I explicitly think about the concepts that underlie the problem.
10. The main skill I got out of Physics 1A was learning how to reason logically about the physical world.
11. To be able to use an equation in a problem (particularly in a problem that I haven't seen before), I need to know more than what each term in the equation represents.
12. The conclusions for 1A Labs in general took more time to complete than I thought they should.
13. I filled out a similar survey for 1AL. (Agree or Disagree only)

#### Short-Answer Questions:

1. What could your TA have done to be more helpful in 1AL?
2. List the three most important physics concepts in 1A.



3. Which concept in 1A did you have the most difficulty with?
4. What did you think the purpose of 1AL was?
5. State the scientific method in your own words.
6. What is the role of uncertainty in the scientific method?
7. How long is the arrow, according to the ruler? What is the *uncertainty* of that measurement?



8. What should an ideal conclusion for Physics Lab class contain?
9. What was the purpose of writing conclusions for 1AL?
10. Why would you want to measure something in lab more than once?

## **Pilot Study (Summer 2012): Post-Instruction**

### **Multiple Choice**

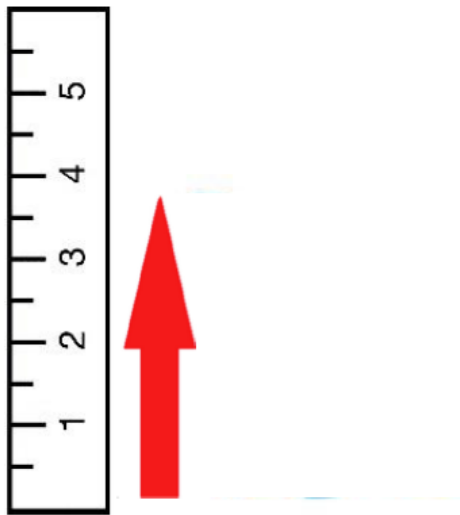
Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree

1. Problem solving in physics basically means matching problems with facts or equations and then substituting values to get a number.
2. In this course, I do not expect to understand equations in an intuitive sense; they just have to be taken as givens.
3. The lab reinforced my understanding of physics concepts from lecture or the book.
4. Physical laws have little relation to what I experience in the real world.
5. In doing a physics problem, if my calculation gives a result that differs significantly from what I expect, I'd have to trust the calculation.
6. The most crucial thing in solving a physics problem is finding the right equation to use.
7. Physics is related to the real world and it sometimes helps to think about the connection, but it is rarely essential for what I have to do in this course.
8. The main skill I get out of this course is learning how to solve physics problems.
9. When I solve most exam or homework problems, I explicitly think about the concepts that underlie the problem.
10. The main skill I get out of 1B is to learn how to reason logically about the physical world.
11. To be able to use an equation in a problem (particularly in a problem that I haven't seen before), I need to know more than what each term in the equation represents.
12. The conclusions for 1B Labs in general took more time to complete than I thought they should.

### **Short-Answer Questions:**

1. What could your TA have done to be more helpful in 1BL?
2. List the three most important physics concepts in the lab.

3. Which concept in the class have you had the most difficulty with?
4. What do you think the purpose of 1BL is?
5. State the scientific method in your own words.
6. What is the role of uncertainty in the scientific method?
7. How tall is the arrow, according to the ruler? What is the *uncertainty* of that measurement?



8. What should an ideal conclusion for Physics Lab class contain?
9. What was the purpose of writing conclusions for 1BL?
10. Why would you want to measure something in lab more than once?

## 1AL (Fall 2012): Pre-Instruction

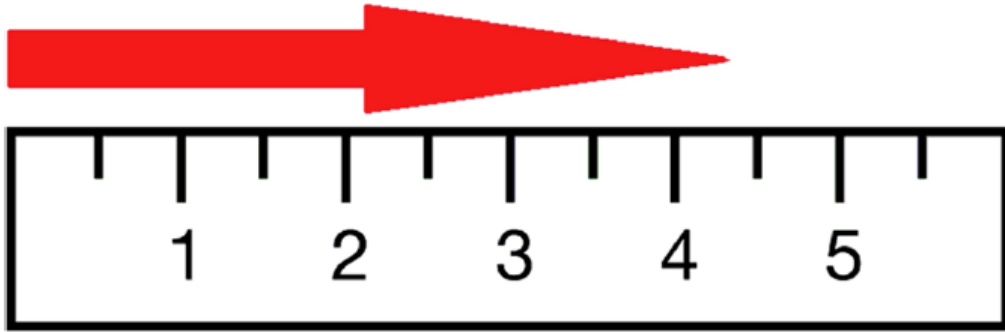
### Multiple Choice

Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree. Please print out this form and answer the questions on it.

1. A significant problem in learning physics is being able to memorize all the information I need to know.
2. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in physics consists of many disconnected topics.
4. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. I cannot learn physics if the teacher does not explain things well in class.
6. In physics, it is important for me to make sense out of formulas before I can use them correctly.
7. I can usually figure out a way to solve physics problems.
8. The subject of physics has little relation to what I experience in the real world.
9. It is possible to explain physics ideas without mathematical formulas.
10. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.
11. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.
12. I have taken a physics class before this one (Agree or Disagree only).

### Short-Answer Questions:

1. What is the role of uncertainty in the scientific method?



Answer the following three questions by examining the picture above

2. How long is the arrow, according to the ruler?

3. What is the *uncertainty* of that measurement?

4. How did you determine the uncertainty of that measurement?

5. Why would you want to measure something in lab more than once?

6. Two groups of students measure the time it takes for a 100 g ball to fall 2.5 meters. Group A obtains the result  $t = 732 \pm 6$  ms, while Group B measures  $t = 741 \pm 7$  ms.

Do the results of the two groups agree? Why / why not?

## 1AL (Fall 2012): Post-Instruction

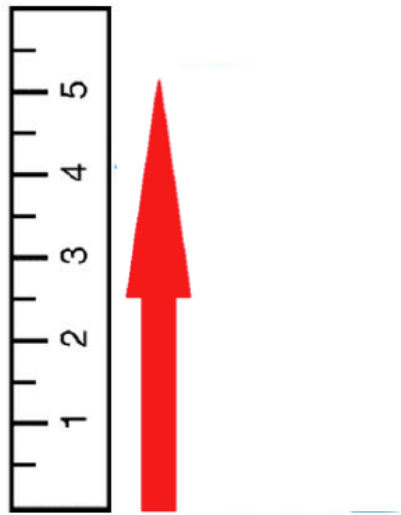
### Multiple Choice

Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree. Please print out this form and answer the questions on it.

1. A significant problem in learning physics is being able to memorize all the information I need to know.
2. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in physics consists of many disconnected topics.
4. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. I cannot learn physics if the teacher does not explain things well in class.
6. In physics, it is important for me to make sense out of formulas before I can use them correctly.
7. I can usually figure out a way to solve physics problems.
8. The subject of physics has little relation to what I experience in the real world.
9. It is possible to explain physics ideas without mathematical formulas.
10. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.
11. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.
12. I have taken a physics class before this one (Agree or Disagree only).
13. The conclusions for 1A Labs in general took more time to complete than I thought they should.

### Short-Answer Questions:

1. What could your TA have done to be more helpful in 1AL?
2. What is the role of uncertainty in the scientific method?



Answer the following three questions by examining the picture above.

2. How tall is the arrow, according to the ruler?
3. What is the *uncertainty* of that measurement?
4. How did you determine the uncertainty of that measurement?
5. Why would you want to measure something in lab more than once?
6. Two groups of students measure the range of a ball that has rolled off a ramp. Group A measures  $x = 86 \pm 3$  cm, while Group B measures  $x = 81 \pm 2$  cm.  
Do the results of the two groups agree? Why / why not?

## 1BL (Winter 2013): Pre-Instruction

### Multiple Choice [Lab 2 Survey]

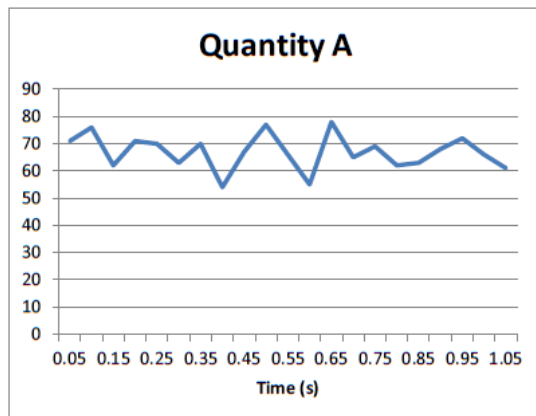
Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree. Please print out this form and answer the questions on it.

1. A significant problem in learning physics is being able to memorize all the information I need to know.
2. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in physics consists of many disconnected topics.
4. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. I cannot learn physics if the teacher does not explain things well in class.
6. In physics, it is important for me to make sense out of formulas before I can use them correctly.
7. I can usually figure out a way to solve physics problems.
8. The subject of physics has little relation to what I experience in the real world.
9. It is possible to explain physics ideas without mathematical formulas.
10. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.
11. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.
12. I took Physics 1AL at UCSD in Fall 2012 (Agree or Disagree only).

### Short-Answer Questions:

1. What is the role of uncertainty in the scientific method?





You use a Logger Pro-like data acquisition system to make the following measurements of Quantity A, as shown in the picture above. Using this data, answer questions 2, 3, and 4:

2. How large is quantity A?

3. What is the uncertainty in this measurement?

4. How would you record this measurement if it were required for a lab conclusion?

5. Why would you want to measure something in lab more than once?

6. A labmate measures the velocity of a glider as  $1.37 \pm 0.08$  m/s. What does that imply about the velocity of the glider?

7. Why do we ask you to measure uncertainties in the lab?

## 1BL (Winter 2013): Post-Instruction

### Multiple Choice [Week 9 Survey]

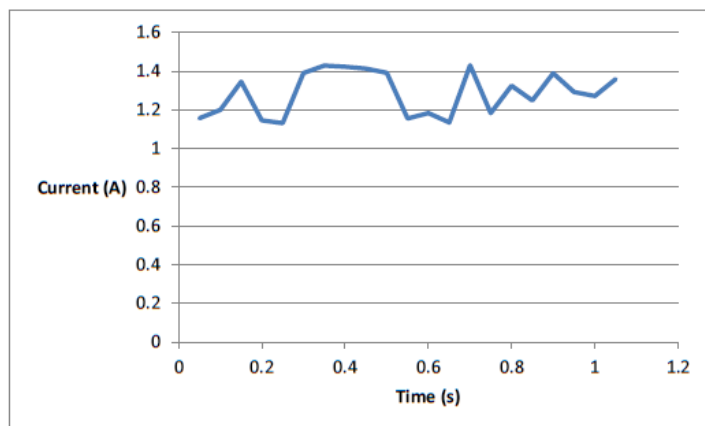
Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree. Please print out this form and answer the questions on it.

1. A significant problem in learning physics is being able to memorize all the information I need to know.
2. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in physics consists of many disconnected topics.
4. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. I cannot learn physics if the teacher does not explain things well in class.
6. In physics, it is important for me to make sense out of formulas before I can use them correctly.
7. I can usually figure out a way to solve physics problems.
8. The subject of physics has little relation to what I experience in the real world.
9. It is possible to explain physics ideas without mathematical formulas.
10. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.
11. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.
12. I took Physics 1AL at UCSD in Fall 2012 (Agree or Disagree only).
13. I turned in a similar survey in Week 2 (Agree or Disagree only).
14. The conclusions for 1A Labs in general took more time to complete than I thought they should.

### Short-Answer Questions:

1. Why do we ask you to measure uncertainties in the lab?

2. What is the role of uncertainty in the scientific method?



You use Logger Pro to measure the current through a resistor. The resistor is assumed to be exactly 20 Ohms. Use the picture above to answer questions 3, 4, 5 and 6:

3. What is the current through the resistor?

4. What is the uncertainty of that current?

5. If your lab conclusion asked you to report the voltage drop across the resistor according to your data, what would you write?

6. A student from another labgroup calculates a voltage of 22.0 V for the voltage across the resistor. Is that result consistent with your data? Explain.

7. Why would you want to measure something in lab more than once?

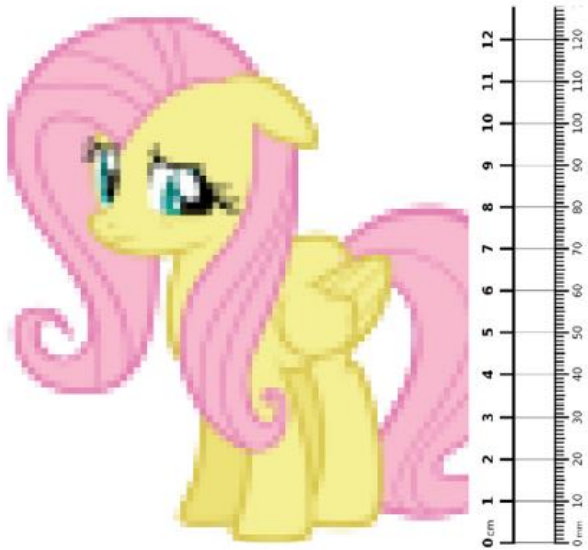
## 1CL (Spring 2013): Pre-Instruction

### Lab 2 Survey:

Answering these survey questions is not mandatory and will have no effect on your grade, but will help us improve future lab activities. Do not write your name on this survey.

### Short-Answer Questions:

1. Why do we ask you to measure uncertainties in the lab?



2. The above picture of Fluttershy is pixelated. Using the ruler on the right, to what precision could you determine her height? (Assume you would be measuring from her bottom foot to the top of her lovely hair.)

3. A lab-mate reports Fluttershy's height as  $124 \text{ mm} \pm 1 \text{ mm}$ . Would you agree or disagree with this statement? Why or why not?

**Multiple Choice Questions:**

**Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree.**

1. A significant problem in learning physics is being able to memorize all the information I need to know.
2. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in physics consists of many disconnected topics.
4. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. I cannot learn physics if the teacher does not explain things well in class.
6. In physics, it is important for me to make sense out of formulas before I can use them correctly.
7. I can usually figure out a way to solve physics problems.
8. The subject of physics has little relation to what I experience in the real world.
9. It is possible to explain physics ideas without mathematical formulas.
10. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.
11. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.
12. I took Physics 1AL at UCSD in Fall 2012 (Agree or Disagree only).
13. I took Physics 1BL at UCSD in Winter 2013 (Agree or Disagree only).

## **1CL (Spring 2013): Post-Instruction**

### **Lab 9 Survey:**

Answering these survey questions is not mandatory and will have no effect on your grade, but will help us improve future lab activities. Do not write your name on this survey.

### **Short-Answer Questions:**

Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree. Then elaborate on your response in the space provided.

1. Did you find conclusion prompts helpful?
  
  
  
  
  
  
  
  
  
  
2. Did you find checkpoint questions helpful?
  
  
  
  
  
  
  
  
  
  
3. Did you find TAs commenting on your graded work helpful?
  
  
  
  
  
  
  
  
  
  
4. Did your understanding of uncertainty change in the three quarters you have taken this class?
  
  
  
  
  
  
  
  
  
  
5. Did you find the pacing of the lecture with respect to the labs helpful? (Please include the name of your professor)
  
  
  
  
  
  
  
  
  
  
6. Please list any other ways the labs could have helped you learn.

**Multiple Choice Questions:**

**Answer these questions on the following scale: Strongly Disagree, Disagree, Neither Disagree nor Agree, Agree, Strongly Agree.**

1. A significant problem in learning physics is being able to memorize all the information I need to know.
2. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in physics consists of many disconnected topics.
4. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. I cannot learn physics if the teacher does not explain things well in class.
6. In physics, it is important for me to make sense out of formulas before I can use them correctly.
7. I can usually figure out a way to solve physics problems.
8. The subject of physics has little relation to what I experience in the real world.
9. It is possible to explain physics ideas without mathematical formulas.
10. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.
11. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.
12. The conclusions for 1CL Labs in general took more time to complete than I thought they should.
13. I took Physics 1AL at UCSD in Fall 2012 (Agree or Disagree only).
14. I took Physics 1BL at UCSD in Winter 2013 (Agree or Disagree only).

## Appendix IV: CLASS Item Data

For each class, the fraction of expert-like responses to each survey item are included both pre- and post-instruction. Significant p-values are bolded. See Chapter 3 for more details.

### 1AL (Fall 2012) CLASS items

N = 47 pre-instruction; N = 188 post-instruction.

CLASS Item	Pre-Inst	Post-Instr	p-val
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.32	0.28	0.28
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.26	0.23	0.32
6. Knowledge in physics consists of many disconnected topics.	0.68	0.64	0.32
8. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.	0.06	0.09	0.27
12. I cannot learn physics if the teacher does not explain things well in class.	0.06	0.19	0.02
24. In physics, it is important for me to make sense out of formulas before I can use them correctly.	0.87	0.81	0.15
34. I can usually figure out a way to solve physics problems.	0.45	0.39	0.25
35. The subject of physics has little relation to what I experience in the real world.	0.79	0.60	0.008
38. It is possible to explain physics ideas without mathematical formulas.	0.49	0.47	0.40
39. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.	0.62	0.67	0.25
42. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.	0.68	0.57	0.08



### 1BL (Winter 2012) CLASS items

N = 112 pre-instruction; N = 147 post-instruction.

CLASS Item	Pre-Inst	Post-Instr	p-val
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.31	0.30	0.41
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.24	0.22	0.33
6. Knowledge in physics consists of many disconnected topics.	0.61	0.58	0.33
8. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.	0.04	0.05	0.36
12. I cannot learn physics if the teacher does not explain things well in class.	0.11	0.14	0.24
24. In physics, it is important for me to make sense out of formulas before I can use them correctly.	0.82	0.79	0.26
34. I can usually figure out a way to solve physics problems.	0.40	0.44	0.28
35. The subject of physics has little relation to what I experience in the real world.	0.53	0.54	0.39
38. It is possible to explain physics ideas without mathematical formulas.	0.50	0.52	0.35
39. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.	0.64	0.66	0.38
42. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.	0.49	0.58	0.08

### 1CL (Spring 2013) CLASS items

N = 330 pre-instruction; N = 330 post-instruction.

CLASS Item	Pre-Inst	Post-Instr	p-val
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.43	0.42	0.38
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.35	0.34	0.42
6. Knowledge in physics consists of many disconnected topics.	0.64	0.60	0.15
8. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.	0.06	0.07	0.27
12. I cannot learn physics if the teacher does not explain things well in class.	0.20	0.21	0.36
24. In physics, it is important for me to make sense out of formulas before I can use them correctly.	0.81	0.81	0.48
34. I can usually figure out a way to solve physics problems.	0.59	0.58	0.38
35. The subject of physics has little relation to what I experience in the real world.	0.56	0.59	0.23
38. It is possible to explain physics ideas without mathematical formulas.	0.51	0.54	0.22
39. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.	0.66	0.65	0.39
42. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.	0.64	0.59	0.06

## 1CL (Summer 2014) CLASS items

Table of items with an expert consensus.

N = 83 pre-instruction; N = 88 post-instruction.

CLASS Item	Pre-Inst	Post-Inst
1. A significant problem in learning physics is being able to memorize all the information I need to know.	0.30	0.34
2. When I am solving a physics problem, I try to decide what would be a reasonable value for the answer.	0.65	0.88
3. I think about the physics I experience in everyday life.	0.29	0.36
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.	0.18	0.25
6. Knowledge in physics consists of many disconnected topics.	0.41	0.47
8. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values.	0.06	0.14
10. There is usually only one correct approach to solving a physics problem.	0.72	0.69
11. I am not satisfied until I understand why something works the way it does.	0.75	0.70
12. I cannot learn physics if the teacher does not explain things well in class.	0.05	0.05
13. I do not expect physics equations to help my understanding of the ideas; they are just for doing calculations.	0.48	0.67
14. I study physics to learn knowledge that will be useful in my life outside of school.	0.27	0.22
15. If I get stuck on a physics problem on my first try, I usually try to figure out a different way that works.	0.72	0.75
16. Nearly everyone is capable of understanding physics if they work at it.	0.61	0.68
17. Understanding physics basically means being able to recall something you've read or been shown.	0.52	0.48

<b>CLASS Item (cont)</b>	<b>Pre-Inst</b>	<b>Post-Instr</b>
18. There could be two different correct values for the answer to a physics problem if I use two different approaches.	0.51	0.39
19. To understand physics I discuss it with friends and other students.	0.75	0.80
20. I do not spend more than five minutes stuck on a physics problem before giving up or seeking help from someone else.	0.69	0.51
21. If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it.	0.40	0.39
22. If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations.	0.23	0.35
23. In doing a physics problem, if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than going back through the problem.	0.76	0.65
24. In physics, it is important for me to make sense out of formulas before I can use them correctly.	0.75	0.81
25. I enjoy solving physics problems.	0.19	0.23
26. In physics, mathematical formulas express meaningful relationships among measurable quantities.	0.73	0.75
27. It is important for the government to approve new scientific ideas before they can be widely accepted.	0.34	0.31
28. Learning physics changes my ideas about how the world works.	0.53	0.56
29. To learn physics, I only need to memorize solutions to sample problems.	0.84	0.76
30. Reasoning skills used to understand physics can be helpful to me in my everyday life.	0.69	0.64
32. Spending a lot of time understanding where formulas come from is a waste of time.	0.66	0.53
34. I can usually figure out a way to solve physics problems.	0.36	0.39
35. The subject of physics has little relation to what I experience in the real world.	0.54	0.55
36. There are times I solve a physics problem more than one way to help my understanding.	0.48	0.49

<b>CLASS Item (cont)</b>	<b>Pre-Inst</b>	<b>Post-Instr</b>
37. To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed.	0.40	0.45
38. It is possible to explain physics ideas without mathematical formulas.	0.60	0.66
39. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.	0.65	0.69
40. If I get stuck on a physics problem, there is no chance I'll figure it out on my own.	0.51	0.47
42. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented.	0.77	0.66

CLASS items without an expert consensus. Agreement and disagreement fractions with the statements are listed.

<b>CLASS Item</b>		<b>Pre-Inst</b>	<b>Post-Inst</b>
4. It is useful for me to do lots and lots of problems when learning physics.	Agree	0.76	0.73
	Disagree	0.06	0.10
7. As physicists learn more, most physics ideas we use today are likely to be proven wrong.	Agree	0.16	0.20
	Disagree	0.33	0.32
9. I find that reading the text in detail is a good way for me to learn physics.	Agree	0.20	0.30
	Disagree	0.54	0.51
33. I find carefully analyzing only a few problems in detail is a good way for me to learn physics.	Agree	0.40	0.42
	Disagree	0.61	0.28
41. It is possible for physicists to carefully perform the same experiment and get two very different results that are both correct.	Agree	0.52	0.53
	Disagree	0.17	0.15

## Appendix V: Data comparison results

### 1AL Fall 2012 Data Range Comparison Rubric

#### YES:

- A) There isn't a significant difference between the group's results
- B) Everything has error, it's impossible to get exactly the same every time
- C) There's a difference of XX between the two group's averages. XX is small, so they pretty much have the same result
- D) Group A's range is from XX-YY, B's from ZZ-AA, so the ranges overlap
- E) There is a possible value that could fit both ranges
- F) Other yes

#### NO:

- G) Their averages are different
- H) There is a significant difference between the two group's results
- I) The difference of X is a large difference compared to the times they're measuring
- J) Group A's range has a width of XX, B's has a width of YY, so they have different results
- K) Group A's range is from XX to YY, B's from ZZ-AA; they only overlap for CC which is not enough
- L) Group A's average of XX doesn't fall within the range for group B (YY-ZZ) and vice versa
- M) There is not a possible value that could fit both ranges
- N) Other no

#### NOT SURE

- O) Group A's range is from XX to YY, B's from ZZ-AA; they only overlap for CC, which I am not sure is enough overlap for them to agree
- P) There is a possible value that could fit both ranges (as in E, but not sure whether this means they agree)
- Q) Other not sure

## 1AL Fall 2012 Pre-Instruction Prompt and Results

6. Two groups of students measure the time it takes for a 100 g ball to fall 2.5 meters. Group A obtains the result  $t = 732 \pm 6$  ms, while Group B measures  $t = 741 \pm 7$  ms.

Do the results of the two groups agree? Why / why not?

45 students produced 53 arguments about the agreement. Fractions below are fractions of the students making each kind of argument.

Yes arguments (N=32)

a	b	c	d	e	f
0.02	0.02	0.02	0.40	0.13	0.11

No arguments (N=14)

g	h	i	j	k	l	m	n
0.02	0.04	0.04	0.00	0.04	0.09	0.02	0.04

Not sure arguments (N=7)

o	p	q
0.13	0.02	0.00

## 1AL Fall 2012 Post-Instruction Prompt and Results

6. Two groups of students measure the range of a ball that has rolled off a ramp. Group A measures  $x = 86 \pm 3$  cm, while Group B measures  $x = 81 \pm 2$  cm.

Do the results of the two groups agree? Why / why not?

158 students produced 208 arguments about the agreement. Fractions below are fractions of the students making each kind of argument.

Yes arguments (N=140)

a	b	c	d	e	f
0.00	0.00	0.01	0.48	0.35	0.04

No arguments (N=14)

g	h	i	j	k	l	m	n
0.02	0.03	0.00	0.03	0.14	0.01	0.01	0.04

Not sure arguments (N=7)

o	p	q
0.11	0.04	0.01



## 1BL Winter 2013 Post-Instruction Prompt and Results

Students estimated a current value from a graph, then calculated a voltage uncertainty from it. Then they answered the following prompt. See Figure 4-6 and the surrounding discussion for more details.

6. A student from another labgroup calculates a voltage of 22.0 V for the voltage across the resistor. Is that result consistent with your data? Explain.

84 students produced 86 arguments about the agreement. Fractions below are fractions of the students making each kind of argument.

Yes arguments (N=43)

a	b	c	d	e	f
0.04	0.00	0.04	0.27	0.01	0.15

No arguments (N=34)

g	h	i	j	k	l	m	n
0.01	0.01	0.00	0.00	0.24	0.00	0.02	0.12

Not sure arguments (N=9)

o	p	q
0.02	0.01	0.07

## References

1. Trumper, R. "The Physics Laboratory – A Historical Overview and Future Perspectives." *Science & Education*, vol. 12, (2003), pp. 645-70.
2. Arons, A. "Guiding insight and inquiry in the introductory physics laboratory." *The Physics Teacher*, vol. 31, 278 (1993). [aapt.scitation.org/doi/pdf/10.1119/1.2343763](http://aapt.scitation.org/doi/pdf/10.1119/1.2343763). Accessed 11 Aug 2018.
3. Serway, R.A., and Jewett Jr., J.W. *Principles of Physics: A Calculus-based Text, Fifth Edition*. Brooks/Cole. 2013.
4. Brownell, S.E., Kloser, M.J., Fukami, T., and Shavelson, R. "Undergraduate Biology Lab Courses: Comparing the Impact of Traditionally Based "Cookbook" and Authentic Research-Based Courses on Student Lab Experiences." *Journal of College Science Teaching*, vol. 4, 4 (2012), pp. 36-45.
5. Lippmann, R.F. *Students' understanding of measurement and uncertainty in the physics laboratory: social construction, underlying concepts, and quantitative analysis*. 2003. University of Maryland, College Park, PhD dissertation.
6. Allie, S., Buffler, A., Campbell, B and Lubben, F. "First-year physics students' perceptions of the quality of experimental measurements." *International Journal of Science Education*, vol. 20, 4, (1998), pp. 447-59.
7. Lubben, F., and Millar, R. "Children's ideas about the reliability of experimental data." *International Journal of Science Education*, vol. 18, 8, (1996), pp. 955-68.
8. Deardorff, D. *Introductory Physics Students' Treatment of Measurement Uncertainty*. 2001. North Carolina State University, PhD dissertation.
9. Pollard, B., Hobbs, R., Stanley, J.T., Dounas-Frazer, D.R., and Lewandowski, H.J. "Impact of an introductory lab on students' understanding of measurement uncertainty." *2017 Physics Education Research Conference Proceedings*, pp. 312-5.
10. Lewandowski, H.J., Hobbs, R., Stanley, J.T., Dounas-Frazer, D.R., and Pollard, B. "Student reasoning about measurement uncertainty in an introductory lab course." *2017 Physics Education Research Conference Proceedings*, pp. 244-7.
11. Majiet, N. and Allie, S. "Student understanding of measurement and uncertainty: probing the mean." *2018 Physics Education Research Conference*. <http://dx.doi.org/10.1119/perc.2018.pr.Majiet>.
12. Abbott, D. *Assessing student understanding of measurement and uncertainty*. 2003. North Carolina State University, PhD dissertation.

13. Day, J., Holmes, N.G., Roll, I., and Bonn, D.A. "Finding evidence of transfer with invention activities: teaching the concept of weighted average." *2013 Physics Education Research Conference Proceedings*, pp. 117-20.
14. Kung, R.L., and Linder, C. "Metacognitive activity in the physics student laboratory: is increased metacognition necessarily better?" *Metacognition Learning*, vol. 2 (2007), pp 41-56.
15. Holmes, N.G. and Bonn, D.A. "Doing science or doing a lab? Engaging students with scientific reasoning during physics lab experiments." *2013 Physics Education Research Conference Proceedings*, pp. 185-88.
16. "Undergraduate Enrollment Statistics." *UC San Diego Institutional Research*, [ir.ucsd.edu/undergrad/stats-data/enrollment/undergrad.html](http://ir.ucsd.edu/undergrad/stats-data/enrollment/undergrad.html). Accessed 19 July 2018.
17. "Best Global Universities for Biology and Biochemistry." *US News and World Report Education*, [www.usnews.com/education/best-global-universities/biology-biochemistry](http://www.usnews.com/education/best-global-universities/biology-biochemistry). Accessed 19 July 2018.
18. Taylor, J.R. *An Introduction to Error Analysis, Second Edition*. University Science Books, 1997.
19. "Bloom's Taxonomy of Measurable Verbs." *Utica College*. [www.utica.edu/academic/Assessment/new/Blooms Taxonomy - Best.pdf](http://www.utica.edu/academic/Assessment/new/Blooms%20Taxonomy%20-%20Best.pdf)
20. Crouch, C.H., and Mazur, E. "Peer Instruction: Ten years of experience and results." *American Journal of Physics*, vol. 69, 9 (2001), pp. 970-7.
21. "PhET: Interactive simulations for science and math." *University of Colorado Boulder*. [phet.colorado.edu](http://phet.colorado.edu). Accessed 26 July 2018.
22. Bertram Gallant, T., Anderson, M.G., and Killoran C. "Academic integrity in a mandatory physics lab: the influence of post-graduate aspirations and grade point averages." *Science Engineering Ethics*, vol. 19, 1, (2013).
23. Kung, R.L., and Linder, C. "University students' ideas about data processing and data comparison in a physics laboratory course." *Nordic Studies in Science Education*, vol. 4, (2006), pp. 40-53.
24. Adams, W.K., et al. "New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Study." *Physical Review Special Topics – Physics Education Research*, vol. 2, 010101 (2006), pp. 1-14.
25. Zwickl, B.M., Finkelstein, N., and Lewandowski, H.J. "Development and Validation of the Colorado Learning Attitudes about Science Study for Experimental Physics." *Proceedings of the Physics Education Research Conference*, 1513, (2013), pp 442-445.

26. Chi, M.T.H., Feltovich, P.J., and Glaser, R. "Categorization and Representation of Physics Problems by Experts and Novices." *Cognitive Science*, vol. 5, pp. 121-52.
27. Adams, W.K., Perkins, K.K., Dubson, M., Finkelstein, N.D. and Wieman, C.E. "The Design and Validation of the Colorado Learning Attitudes about Science Study." *Physics Education Research Conference, American Institute of Physics Conference Proceedings*, vol. 790 (2005), pp.45-8.
28. Gire, E. *Between the poles: locating physics majors in the expert-novice continuum*. 2007. University of California, San Diego, PhD dissertation.
29. Perkins, K.K., Gratny, M.M., Adams, W.K., Finkelstein, N.D. and Wieman, C.E. "Towards characterizing the relationship between students' interest in and their beliefs about physics." *American Institute of Physics Conference Proceedings*, vol. 818, 137 (2006).
30. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences: Second Edition*. Lawrence Erlbaum Associates. 1998
31. Heitin, L. "2 in 5 High Schools Don't Offer Physics, Analysis Finds." *Education Week*, vol. 36, 1, <https://www.edweek.org/ew/articles/2016/08/24/2-in-5-high-schools-dont-offer.html>.
32. Wesp, R., Cichello, P., Gracia, E.B. and Davis, K. "Observing and engaging in purposeful actions with objects influences estimates of their size." *Perception & Psychophysics*, vol. 66, 8, (2004), pp. 1261-7.
33. Cohen, J., Vanderplas, J.M. and White, W.J. "Effect of Viewing Angle and Parallax upon Accuracy of Reading Quantitative Scales." *The Journal of Applied Psychology*, vol. 37, 6, (1953), pp. 482-8.
34. Kullback, S. and Leibler, R.A. "On Information and Sufficiency." *The Annals of Mathematical Statistics*, vol. 22, 1, (1951), pp. 79-86.
35. Johnson, D.H. and Sinanović, S. "Symmetrizing the Kullback-Leiber distance." Computer and Information Technology Institute, Rice University, (2001).
36. Kung, R.L. "Teaching the concepts of measurement: An example of a concept-based laboratory course." *American Journal of Physics*, vol. 73, 8, (2005), pp.771-7.
37. Kung, R.L. and Linder, C. "Metacognitive activity in the physics student laboratory: is increased metacognition necessarily better?" *Metacognition Learning*, vol. 2 (2007), pp.41-56.
38. "Goals of the Physics Major (USLI)." *University of California, Berkeley*. <http://physics.berkeley.edu/academics/undergraduate-degree/the-major-and-minor-program/goals-of-the-physics-major-usli>. Accessed 2 August 2018.

39. “Physics Bachelors: One Year After Degree.” *American Institute of Physics Statistical Research Center*. 2016. [www.aip.org/statistics/reports/physics-bachelors-one-year-after-degree](http://www.aip.org/statistics/reports/physics-bachelors-one-year-after-degree). Accessed 2 Aug 2018.
40. “UC San Diego – WASC Exhibit 7.1: Inventory of Educational Effectiveness Indicators.” *Department of Biology, University of California, San Diego*. 2014.
41. “General Biology Sample Plan.” *Biology Undergraduate Student and Instructional Services, University of California, San Diego*. 2014.
42. Bernoilles, S. Private interview. 2014.
43. Weinhausen, G. Private interview. 19 May 2014.
44. Epp, C.D. and Bull, S. “Uncertainty Representation in Visualization of Learning Analytics for Learners: Current Approaches and Opportunities.” *Institute of Electrical and Electronics Engineers Transactions on Learning Technologies*, vol. 8, 3, (July-September 2015), pp. 242-60.
45. Bransford, J.D. and Cocking, R.R., eds. *How People Learn: Brain, Mind, Experience and School: Expanded Edition*. National Academy Press, 2000.
46. Perkins, K.K. and Gratny, M. “Who Becomes a Physics Major? A Long-term Longitudinal Study Examining the Roles of Pre-college Beliefs about Physics and Learning Physics, Interest, and Academic Achievement.” *Physics Education Research Conference, American Institute of Physics*, CP1289 (2010), pp. 253-6.
47. Hazari, Z., Sonnert, G., Sadler, P.M. and Shanahan, M.C. “Connecting High School Physics Experiences, Outcome Expectations, Physics Identity and Physics Career Choice: A Gender Study.” *Journal of Research in Science Teaching*, vol. 47, 8, (2010), pp. 978-1003.
48. Ivie, R., White, S., Garrett, A. and Anderson, G. “Women among Physics and Astronomy faculty: Results from the 2010 Survey of Physics Degree-Granting Departments.” *American Institute of Physics Statistical Research Center*. August 2013. [www.aip.org/statistics/reports/women-among-physics-astronomy-faculty](http://www.aip.org/statistics/reports/women-among-physics-astronomy-faculty). Accessed 10 Aug 2018.
49. Ivie, R., Anderson, G., and White, S. “African Americans and Hispanics among Physics and Astronomy faculty: Results from the 2012 Survey of Physics and Astronomy Degree-Granting Departments.” *American Institute of Physics Statistical Research Center*. July 2014. [www.aip.org/statistics/reports/african-americans-hispanics-among-physics-astronomy-faculty-0](http://www.aip.org/statistics/reports/african-americans-hispanics-among-physics-astronomy-faculty-0). Accessed 10 Aug 2018.
50. Harrison, C., and Tanner, K.D. “Language Matters: Considering Microaggressions in Science.” *Cell Biology Education – Life Sciences Education*, vol. 17, fe4 (Spring 2018), pp. 1-8.

51. Wieman, C.E., and Adams, W.K. “On the Proper Use of Statistical Analyses; a Comment on “Evaluation of Colorado Learning Attitudes about Science Survey” by Douglas et al.” [arxiv.org/abs/1501.03257](https://arxiv.org/abs/1501.03257). Accessed 10 Aug 2018.
52. Perkins, K.K., Gratny, M.M., Adams, W.K., Finkelstein, N.D. and Wieman, C.E. “Towards characterizing the relationship between students’ interest in and their beliefs about physics.” *American Institute of Physics Conference Proceedings*, vol. 818, 137 (2006).