ABSTRACT

ABBOTT, DAVID SCOT. Assessing student understanding of measurement and uncertainty. (Under the direction of Robert J. Beichner).

A test to assess student understanding of measurement and uncertainty has been developed and administered to more than 500 students at two large research universities. The aim is two-fold: 1) to assess what students learn in the first semester of introductory physics labs and 2) to uncover patterns in student reasoning and practice. The forty minute, eleven item test focuses on direct measurement and student attitudes toward multiple measurements. After one revision cycle using think-aloud interviews, the test was administered to students to three groups: students enrolled in traditional laboratory lab sections of first semester physics at North Carolina State University (NCSU), students in an experimental (SCALE-UP) section of first semester physics at NCSU, and students in first semester physics at the University of North Carolina at Chapel Hill. The results were analyzed using a mixture of qualitative and quantitative methods. In the traditional NCSU labs, where students receive no instruction in uncertainty and measurement, students show no improvement on any of the areas examined by the test. In SCALE-UP and at UNC, students show statistically significant gains in most areas of the test. Gains on specific test items in SCALE-UP and at UNC correspond to areas of instructional emphasis. Test items were grouped into four main aspects of performance: “point/set” reasoning, meaning of spread, ruler reading and “stacking.” Student performance on the pretest was examined to identify links between these aspects. Items within each aspect are correlated to one another, sometimes quite strongly, but items from different aspects rarely show statistically significant correlation. Taken together, these results suggest that student difficulties may not be linked to a single underlying cause. The study shows that current instruction techniques improve student understanding, but that many students exit the introductory physics lab course without appreciation or coherent understanding for the concept of measurement uncertainty.
Assessing student understanding
of measurement and uncertainty

by

David Scot Abbott

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

Physics

Raleigh, NC

2003

Approved by:

Dr. Robert Beichner
(Chair of Advisory Committee)

Dr. John Hubisz

Dr. J. Richard Mowat

Dr. John Park
DEDICATION

I dedicate this dissertation to my grandfather, Werner Dunkel. One of his dreams was to have one of his three grandchildren complete a Ph.D. When he died on January 13, 2002, he knew that my brother, Ed, and I were well on our way with successful graduate school careers in Ph.D. programs at major universities. This completed dissertation is for you, Grandfather, as well as for me. Thanks for all your support and encouragement!
BIOGRAPHY

I arrived at North Carolina State University in August 1997. After teaching at community colleges for five years near my home in Wilmington, DE, I decided to return to graduate school in physics education. Upon arrival, I viewed the journey with anticipation and trepidation, since my first experience in graduate school had been painful and largely unsuccessful.

The second time around was much more pleasant than the first. Academically and personally, the time at State has been outstanding. I passed (and even understood) courses that I had failed elsewhere. I designed, undertook and completed several research projects. I published articles for the first time. The faculty is largely to blame. Many professors, especially Bob Beichner, helped me in tangible and intangible ways. I met many new friends during my time at State, academic, musical and personal. I recorded my first CD while I was here. Two other full length CD’s followed. A fourth is in production. I met people who will be lifelong friends and fell in love with the woman who will be my wife.

These have been a magical six years. Now I move on to Dartmouth College to put what I have learned about student learning in the lab into practice. I hope that the next six years hold at least as much magic, although it’s tough to imagine how that’s possible!
ACKNOWLEDGEMENTS

I would like to thank the Physics Departments at North Carolina State University and the University of North Carolina at Chapel Hill for allowing this test to be administered in their teaching laboratories. Special thanks go to Jim Chilton and Duane Deardorff who oversee the teaching labs at these two institutions. Research into teaching and learning is impossible without the cooperation of classroom teachers, including the instructors and teaching assistants who administered the test used in this study.

The National Science Foundation provided me with valuable financial support. The NSF fellowship funded through a grant written by Bob Beichner and others paid for the first five years of my education at North Carolina State University. Further assistance was provided indirectly through NSF’s support of the SCALE-UP project.

My advisor, Bob Beichner, has been an invaluable resource. Throughout my graduate career, he has given kind and thoughtful support. Dr. Beichner has taught me many valuable things about teaching, learning, research and life that I may not have otherwise learned. My colleagues in the Physics Education Group have also provided valuable insights. Conversations with Duane Deardorff were instrumental in the research presented in this thesis. Conversations with other graduate students and postdocs, including Scott Bonham, Jeff Saul, Melissa Dancy, Rhett Allain, Jeanne Morse, Matt Kohlmyer, and Sejung Kim were also helpful. Thanks also to Larry Martin and Peg Gjersten, for many valuable contributions.

Thanks, above all, to those who helped me stay sane. Musical friends and collaborators were instrumental (and vocal): Lou Matza, Jim Allen, Ed Hurt, Dottie Northup, Ian Schreier, Carla Burgess, Jerry Kee, Thomas Wilson, Danny Moses, Brent Wissick and all the members of the UNC viol consort. Special thanks go to family (present and future) and friends who make all of life possible and joyful: Dad, Ed, Becky, Gretchen (my fiancee), Lou, Ed Hurt, and Felipe.
# TABLE OF CONTENTS

**LIST OF FIGURES** ........................................................................................................... vii  
**LIST OF TABLES** ........................................................................................................... viii  

## CHAPTER 1: INTRODUCTION ..................................................................................1  
1.1 Motivation for the research ..................................................................................1  
1.2 Research Objectives/Questions ..........................................................................2  
1.3 Relationship to existing research .........................................................................2  
1.4 The role of the lab ...............................................................................................4  

## CHAPTER 2: LITERATURE REVIEW ......................................................................5  
2.1 Expert Understanding of Measurement and Uncertainty .................................5  
   2.1.1 ISO: Terminology and the nature of measurement and uncertainty ..............5  
   2.1.2 What do physicists feel students should know about uncertainty? ..............9  
2.2 Previous Studies on Student Understanding .......................................................9  
   2.2.1 Séré et al.: Studying students during lab work ...........................................9  
   2.2.2 Lubben et al.: Multiple measurements and models of student thinking .......11  
   2.2.3 Deardorff: Treatment of uncertainty by students .......................................15  
   2.2.4 Other efforts ................................................................................................17  
   2.2.5 Synthesis of previous research ....................................................................18  

## CHAPTER 3: METHODOLOGY ...............................................................................21  
3.1 Sample Description ..............................................................................................21  
3.2 Description of Lab Courses at UNC and NCSU ...............................................25  
3.3 Test Construction ................................................................................................27  
   3.3.1 Focus of the test ............................................................................................27  
   3.3.2 Repeated measurements, reading scales and stacking ..............................28  
   3.3.3 Item details ..................................................................................................29  
3.4 Think-Aloud Interviews .......................................................................................40  
3.5 Analysis Methods ................................................................................................42  
3.6 Reliability and Validity .......................................................................................48  
3.7 Summary .............................................................................................................49
CHAPTER 4: ANALYSIS AND FINDINGS .............................................................50

4.1 COMPARISON OF PRETEST AND POSTTEST .........................................................50
  4.1.1 Questions 1 and 2: Purpose of repeat measurements .........................................50
  4.1.2 Question 3: Inclusion or exclusion of an outlier ....................................................55
  4.1.3 Question 4: Spread as a measure of data quality ..................................................57
  4.1.4 Question 5: Assessing agreement between sets of repeat measurements ............58
  4.1.5 Question 6: Reading a ruler .................................................................................61
  4.1.6 Question 7: Stacking ............................................................................................62
  4.1.7 Analyzing the responses to items 8-11 ..................................................................65
  4.1.8 Question 8: Removing an outlier to calculate a mean ..........................................66
  4.1.9 Question 9: Measuring the diameter of a quarter ...............................................70
  4.1.10 Question 10: Measuring the mass of a paper clip ...............................................75
  4.1.11 Question 11: Timing the fall of a coffee filter ....................................................78
  4.1.12 Section summary: Pretest/posttest differences ..................................................81

4.2 ARE DIFFERENT MODES OF THINKING RELATED? ...........................................81
  4.2.1 Internal consistency: Reading a ruler ..................................................................82
  4.2.2 Internal consistency: Stacking .............................................................................86
  4.2.3 Internal consistency: Outlier exclusion ...............................................................87
  4.2.4 Internal consistency: Meaning of spread .............................................................88
  4.2.5 Is point/set reasoning related to other modes of thinking? .................................90
  4.2.6 Is ruler reading practice related to stacking? .......................................................90
  4.2.7 Other related modes of thinking .........................................................................91
  4.2.8 Section summary: Are different modes of thinking related? .............................91

CHAPTER 5: DISCUSSION ..........................................................................................92

5.1 OVERVIEW ............................................................................................................92
5.2 IMPACT OF INSTRUCTION ON STUDENT UNDERSTANDING ..........................92
5.3 RELATIONSHIPS AMONG VARIOUS MODES OF THINKING ..............................94
5.4 IMPLICATIONS FOR INSTRUCTION .....................................................................96
5.5 IMPLICATIONS AND DIRECTIONS FOR FUTURE RESEARCH .............................97

BIBLIOGRAPHY ........................................................................................................99

APPENDICES ..........................................................................................................101

APPENDIX A: LAB INSTRUCTIONS AND ASSIGNMENTS FOR STUDENTS IN SCALE-UP 102
  A.1 Pennies (Lab 1) .....................................................................................................103
  A.2 Spring Lab (Lab 2) ...............................................................................................114
  A.3 Friction (Lab 3) ....................................................................................................124
  A.4 Rolling Objects (Lab 4) ........................................................................................131
  A.5 Pendulum (Lab 5) ................................................................................................146
  A.6 Simple harmonic motion (Lab 6) ..........................................................................156

APPENDIX B: UNCERTAINTY GUIDE FOR STUDENTS IN SCALE-UP ..........................162
LIST OF FIGURES

Figure 2-1. Ball and ramp context for questions about multiple measurements.................................13
Figure 3-1. Experimental context for questions about multiple measurements......................................29
Figure 3-2. Visualization of the data presented in question #3.................................................................32
Figure 3-3. Visualization of the data presented in question #5.................................................................33
Figure 3-4. Mosaic plot of the fictitious data shown in Table 3-6..............................................................46
Figure 3-5. Mosaic plot of fictitious data relating car and snowblower ownership....................................47
Figure 4-1. Classifying reasoning as “point” or “set.” ...............................................................................53
Figure 4-2. Pretest/posttest comparison for questions 1 and 2.................................................................55
Figure 4-3. Pretest/posttest comparison for question 3......................................................................... .....56
Figure 4-4. Pretest/posttest comparison for question 4......................................................................... .....57
Figure 4-5. Pretest/posttest comparison for question 5: Do results agree?.................................................58
Figure 4-6. Pretest/posttest comparison: How students assess agreement on question 5.........................60
Figure 4-7. Pretest/post comparison for question 6...................................................................................61
Figure 4-8. Pretest/posttest comparison of stacking in question 7..............................................................63
Figure 4-9. Pretest/posttest comparison of correct answers to question 7..................................................64
Figure 4-10. How stacking influences student answers to question 7.....................................................64
Figure 4-11. Pretest/posttest comparison of outlier exclusion on question 8.............................................67
Figure 4-12. Pretest/posttest comparison for question 8: Correct mass?.....................................................68
Figure 4-13. Pretest/posttest comparison for question 9..........................................................................72
Figure 4-14. Student measurements: histogram #1.................................................................................73
Figure 4-15. Student measurements: histogram #2.................................................................................73
Figure 4-16. Histogram of the last digit in diameter value.................................................................74
Figure 4-17. Massing a paper clip: Do students stack?.............................................................................76
Figure 4-18. Massing a single paper clip.................................................................................................77
Figure 4-19. Pretest/posttest comparison: Use of standard deviation......................................................79
Figure 4-20. Novice methods for estimating uncertainty............................................................................80
Figure 4-21. Ruler reading practice on questions 6 and 7.......................................................................84
Figure 4-22. Ruler reading practice on questions 6 and 9.......................................................................85
Figure 4-23. Student ruler reading practice on questions 7 and 9............................................................85
Figure 4-24. Stacking on questions 7 and 10............................................................................................87
Figure 4-25. Outlier exclusion on questions 3 and 8...............................................................................88
Figure 4-26. Advanced set reasoning on questions 4 and 5.................................................................89
LIST OF TABLES

Table 3-1. SAT math and verbal scores at UNC and NCSU ................................................................. 22
Table 3-2. Composite SAT scores at UNC and NCSU ................................................................. 22
Table 3-3. High school rank of incoming freshmen at UNC and NCSU .......................................... 22
Table 3-4. Gender balance at UNC and NCSU ............................................................................. 22
Table 3-5. Z-values and confidence intervals .......................................................................................... 43
Table 3-4. Fictitious data for example mosaic plot ........................................................................... 46
Table 4-1. Student reasoning on questions 1 and 2 .............................................................................. 51
Table 4-2. Response categories for anomaly question ........................................................................ 56
Table 4-4. Popular student responses for the mass of the marble in question 8 ......................... 67
Table 4-5. Inferring ruler reading practice from student’s letter choice for question 6 ............. 83
Table 4-6. Inferring ruler reading practice from question 7 .............................................................. 83
Chapter 1: Introduction

1.1 Motivation for the research

Measurements are central to quantitative endeavors. In science, data are used to evaluate hypotheses. Engineers use tables of measurements to make design choices. Economists and other social scientists use measurements to make predictions. Throughout society, important decisions are made based on the result of measurement. The quality of all decisions based on data depends vitally on two components: the quality of the measurements and reasonable estimates for the uncertainty in those measurements and numbers derived from them. Science educators have long recognized the important role that measurements and measurement uncertainty play in science, and teaching objectives referring to measurement and uncertainty have been included in documents outlining national teaching standards (2061 1993; Teachers 1997).

The undergraduate physics laboratory provides an opportunity for students to learn about measurement and uncertainty. In most introductory physics laboratories, students make measurements of quantities that are typically numerical and continuous. These measurements are often used to calculate other quantities in order to test and formulate ideas about the physical world. Questions about the reliability of the numerical results and agreement between two results would seem to rise naturally from such a curriculum and drive students to form a better understanding of measurements and measurement uncertainty. However, recent research (Sere, Journeaux et al. 1993; Lubben and Millar 1996; Allie, Buffler et al. 1998; Coelho and Sere 1998; Deardorff 2001; Lubben, Campbell et al. 2001) has found that high school, college and even some graduate students have noticeable difficulty with making valid, reliable measurements and properly assessing the reliability and validity of their measurements.

The plan for this study is to develop a test to assess what students learn about measurement and uncertainty and use that test to compare what students learn in several undergraduate physics laboratory courses offered at North Carolina State University (PY205, PY211) and at University of North Carolina (Physics24 and Physics26).
1.2 **Research Objectives/Questions**

This project has two main objectives. The first is to investigate the impact of instruction on student understanding of measurement and uncertainty. The second is to identify patterns of student reasoning about measurement practices and measurement uncertainty. Two main research questions drive the study:

- How does student understanding of measurement change over one semester from participating in an introductory physics course with a laboratory? How does this compare across different introductory physics laboratory curricula?
- What patterns among student practices and reasoning exist across various measurement tasks prior to instruction?

Answering these questions may help to identify elements of successful instruction and provide insight for curriculum development.

1.3 **Relationship to existing research**

Science educators have studied the role of the student laboratory in student learning for almost two hundred years. (Lazarowitz and Tamir 1994) One of the most commonly stated aims of the student science laboratory is to teach students the process of science. Over the last few decades, science education researchers and cognitive scientists have been working to identify the features of the ability to ‘do science’ and how students acquire this ability. Researchers have investigated various aspects of this procedural knowledge, including student understandings of:

- control of variables (Lawson 1985),
- the relationship between scientific ideas and evidence for those ideas (Kuhn, Amsel et al. 1988),
- procedures for determining the relationship between variables (Schauble 1996),
- defining variables (Duggan, Johnson et al. 1996), and
- inherent uncertainty of empirical data (Lubben and Millar 1996).

The cited sources represent only a fraction of the work that has been done. The study of the nature of what it means to ‘do science’ and ‘know science’ is a fundamental feature of the study of performance assessment.
In the newer field of physics education research, physicists, cognitive scientists and others have studied conceptual difficulties that students have in various content areas within physics, from introductory kinematics to quantum mechanics. (Redish and McDermott 1999) These studies have led the development of highly effective curricular materials as well as tests for assessing student understanding and evaluating instruction. Physics education research, however, has focused predominantly on traditional physics content, and largely ignored procedural knowledge associated with scientific research and laboratory work. The present project seeks to apply the practices and tradition of physics education research to the study of procedural knowledge.

Despite the importance of measurement to scientific endeavor, it is only in the last ten years that researchers in science education and physics education have begun to study students’ understanding of measurement and uncertainty. An extensive literature search uncovered only six major studies (Sere, Journeaux et al. 1993; Lubben and Millar 1996; Allie, Buffler et al. 1998; Coelho and Sere 1998; Deardorff 2001; Lubben, Campbell et al. 2001). Only one (Deardorff 2001) of these was conducted in the United States. The main result of these studies is a long list of specific student difficulties associated with measurement uncertainty. Some themes have emerged, but more research needs to be done to more completely understand student reasoning. The impact of instruction on student understanding of measurement also needs more study. Two of the studies cited above (Lubben and Millar 1996; Deardorff 2001) show that older students with more laboratory experience generally perform better than younger, less experienced students, but the effect of instruction has never been studied directly.

The development of a valid and reliable assessment instrument would contribute to the research in two ways. Such a test would allow institutions and researchers alike to evaluate current instructional practices for teaching measurement and uncertainty and serve to guide development of new curriculum. When administered to a large number of students, such a test would also provide a rich database of student responses for discovering patterns in student reasoning.
1.4 The Role of the Lab

It is likely that most of what students learn in physics about measurement and uncertainty is due to laboratory instruction. Measurement and uncertainty is rarely taught in the lecture part of the introductory physics course. Evidence of this can be found in top selling introductory textbooks (Giancoli 1998; Serway and Beichner 2000; Cutnell and Johnson 2001). Apart from a cursory treatment of significant figures in Chapter 1 or in an appendix, measurement uncertainty is not addressed. In direct contrast, physics laboratory manuals usually include reference sections on uncertainty analysis that cover topics such as statistics of repeated measurements, propagation of measurement uncertainty in calculations, significant figures, sources of error, systematic and random errors. Some lab manuals also include activities designed to directly address measurement and uncertainty. Comparing student learning in several lab programs may help to identify key elements of lab instruction which are especially effective at improving student understanding of measurement and uncertainty.
2.1 Expert Understanding of Measurement and Uncertainty

Before attempting to construct a test to assess student understanding of measurement, it is necessary to have some picture of expert understanding of the topic. Such a picture cannot easily be constructed from textbooks. In a recent survey of written materials and measurement experts from metrology, Deardorff found noticeable discrepancies, even among experts and popular error analysis texts (Bevington and Robinson 1992; Baird 1995; Taylor 1997), concerning use of terminology, standards for reporting and estimating uncertainty, and (implicitly) criteria for evaluating agreement between two measured quantities. However, a definitive source does exist. In 1993, the group of international experts organized by the International Standards Organization published two documents designed to eliminate inconsistencies in practice and terminology:


The ISO guidelines do not represent new thinking about measurement, but rather a refinement and codification of the similar, but nonuniform, language and practices described in many sources. For this reason, I choose to adopt the ISO standards as the authoritative source for expert opinion on measurement uncertainty.

2.1.1 ISO: Terminology and the nature of measurement and uncertainty

This section provides a brief description of key definitions from VIM. This serves to clarify the meanings of basic metrology terms as they will be used throughout this dissertation. A brief analysis of these terms will also identify key elements of expert thinking concerning uncertainty and the general nature of measurement.
Central to expert understanding of measurement is the acceptance of variability and uncertainty, as shown by the definition of *true value* in VIM:

**true value (of a quantity)** [VIM 1.19] – value consistent with the definition of a given particular quantity. A true value by nature is indeterminate; this is a value that would be obtained by a perfect measurement (ISO, p. 32).

*Note:* The indefinite article "a," rather than the definite article "the," is used in conjunction with "true value" because there may be many values consistent with the definition of a given particular quantity (ISO, p. 32).

Because a perfect measurement is not possible, even in principle, estimation of the uncertainty of measurements is an essential part of measurement. This is reflected most directly in the definition of *result of a measurement*.

**result of a measurement** [VIM 3.1] – value attributed to a measurand, obtained by measurement. A complete statement of the result of a measurement includes information about the uncertainty of measurement (ISO, p. 33).

Even more compelling evidence of the importance placed on estimating uncertainty by experts is the existence and prominence of the concept of uncertainty itself, as evidenced by the title of one of the two ISO documents.

Because the concept of uncertainty is not typically taught, it is important to carefully examine the definition of *uncertainty* in VIM:

**uncertainty (of measurement)** [VIM 3.9] – parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand. The uncertainty generally includes many components which may be evaluated from experimental standard deviations based on repeated observations (Type A evaluation) or by standard deviations evaluated from assumed probability distributions based on experience or other information (Type B evaluation). The term uncertainty is preferred over measurement error because the latter can never be known (ISO, p. 34).

**standard uncertainty**, \(u_i\) – the uncertainty of the result of a measurement expressed as a standard deviation (ISO, p. 3).

Uncertainty is a numerical estimate of the quality of a measurement. This numerical estimate is made without reference to a true value of the measurand and is based on statistics.
The ISO publications clearly distinguish **uncertainty** from **error**. Error estimates, in direct contrast to uncertainty estimates, are linked to a true value of the measurand:

**error (of measurement)** [VIM 3.10] – result of a measurement minus a true value of the measurand (which is never known exactly); sometimes referred to as the "absolute error" to distinguish from "relative error" (ISO, p. 34).

**random error** [VIM 3.13] – result of a measurement minus the mean that would result from an infinite number of measurements of the same measurand carried out under repeatable conditions (ISO, p. 34).

**systematic error** [VIM 3.14] – mean that would result from an infinite number of measurements of the same measurand carried out under repeatability conditions minus a true value of the measurand; error minus random error (ISO, p. 34).

Error and uncertainty are often used interchangeably by experts, including at least one leading text. Many texts distinguish between random error and systematic error, but fail to distinguish between a priori estimates of the reliability of results (i.e. uncertainty) and post posteriori estimates of the correctness of results (i.e. error). It is interesting to note that, while the ISO Guide distinguishes between random and systematic errors, the Guide advocates that the uncertainties arising from the two different sources be treated identically.

According to its definition given above, uncertainty is statistical in nature. The ISO Guide describes two different ways of estimating uncertainty in measured quantities, termed Type A and Type B. Both are based on statistics:

Thus a Type A standard uncertainty is obtained from a **probability density function** (C.2.5) derived from an observed **frequency distribution** (C.2.18), while a Type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur [often called subjective **probability** (C.2.1)]. Both approaches employ recognized interpretations of probability. (ISO, p.6)

Uncertainty at its heart is a confidence interval that requires statistical interpretation. Complete reporting of uncertainty must include information about the confidence level associated with the stated interval. Stated another way, what a measurement of (53 cm ± 4 cm) means depends crucially on the meaning of the uncertainty value. Is this a 99% confidence interval or a 67% confidence interval? Likewise, rigorous evaluation of the equality of two results requires probabilistic interpretation. The agreement of two results depends on the two results, their uncertainties and the level of confidence set for agreement.
It should be noted that the definition of standard uncertainty explicitly sets a scale for the size for
“reasonable” uncertainty estimates. This choice is deliberate:

E.1.1 This Guide presents a widely applicable method for evaluating and expressing uncertainty in
measurement. It provides a realistic, rather than “safe” value of the uncertainty based on the concept
that there is no inherent difference between an uncertainty component arising from a random effect
and one arising from a correction due to a systematic effect (see 3.2.2 and 3.2.3). The method stands,
therefore, in direct contrast to certain older methods that have the following two ideas in common.

E.1.2 The first idea is that the uncertainty reported should be “safe” or “conservative,” meaning that
it must never err on the side of being too small. In fact, because the evaluation of the uncertainty of a
measurement result is problematic, it was often made deliberately large. (ISO p.45)

Confidence intervals that are too wide do not effectively convey the limitations of the measurement.

The ISO guides contain much more detail than has been presented here. In particular, the GUM sets
explicit guidelines for the estimation of measurement uncertainty, gives examples showing how uncertainties
in measurements and results are estimated, and presents other information well beyond the scope of first year
undergraduate courses. It should be mentioned that the ISO guidelines do not address several topics associated
with measurement and uncertainty that are often taught to undergraduates, including significant figures. These
omissions are not critical, however, since proper practice can often be inferred from the ISO publications
through applications of principles and concepts from the ISO guides.
2.1.2 What do physicists feel students should know about uncertainty?

A recent survey (Deardorff 2001) of 26 physics graduate students and professors asked respondents to answer the following question:

What do you think are the most important concepts or skills students should learn about measurement uncertainty and error analysis? (Deardorff, p.25)

The main concepts and skills these physics teachers identified (as categorized by Deardorff) are:

- All measured values have uncertainty
- Uncertainties must be estimated and clearly reported
- Reporting proper number of significant figures
- Propagation of uncertainty
- Identify and classify sources of uncertainty

The responses generally indicate that these physics teachers felt that students should understand basic concepts.

2.2 Previous studies on student understanding

Despite its importance, study of student understanding of measurement is still in its infancy. Only five published research journal articles (Sere, Journeaux et al. 1993; Lubben and Millar 1996; Allie, Buffler et al. 1998; Coelho and Sere 1998; Lubben, Campbell et al. 2001), one doctoral dissertation (Deardorff 2001) and a handful of shorter conference proceeding papers could be found in the literature. This small body of work reports the findings of essentially three groups of researchers. This body of research will be examined in detail and then reviewed.

2.2.1 Sére et al.: Studying students during lab work

In the earliest work on student understanding of measurement and uncertainty, Sére and others examine student practices and reasoning about measurement and uncertainty. In 1993, Sére, Larcher and Journeaux (Sere, Journeaux et al. 1993) report a qualitative study conducted at the University of Paris. After receiving lecture instruction on measurement and uncertainty analysis, twenty 2nd semester students conducted experiments in optics and electricity. The researchers analyze the audiotape of the lab work in optics and the
final test results from the students. The main topics of lecture instruction were terminology, estimation of uncertainty from a single measurement, propagation of uncertainty and statistical analysis of multiple measurements. Special instructional effort was given to the advantages of expressing a measurement as a number with a confidence interval and calculation of confidence interval. Despite this instruction, the researchers report substantial conceptual difficulties manifested by their students. The most serious are

- Some students fail to distinguish between uncertainty determined from a single measurement and uncertainty determined from repeated measurements.
- Some students establish a hierarchy in a series of repeated measurements, often viewing later measurements as only providing confirmation for earlier readings.
- Some students fail to distinguish between uncertainty and error (Séré uses the terms precision and accuracy).

The first two difficulties illustrate a profound lack of understanding concerning the purpose behind making multiple measurements. The researchers also note that students do not spontaneously evaluate the uncertainty of the experimental results. As noted by the authors, this may be a reflection of the nature of the lab tasks performed by the students in this study. In both experiments, the product of measurement and calculation was a characteristic of an object (i.e. the focal length of a lens or a value for resistivity of a wire). The authors suggest that experiments linked to testing of models might provide students with better motivation for evaluating the uncertainty in measurements. The implication for teaching is that inquiry-oriented labs may the development of expert-like views of measurement.

In 1998, Coelho and Séré (Coelho and Sere 1998) report findings from some of the data originally reported in Coelho’s 1993 doctoral dissertation. Individual and small group interviews were conducted with 21 French secondary students as they took and interpreted measurements during an experiment in a clinical setting. The students observed as the interviewer ran an air puck with a spark timer over a level surface. Students measured distances between marks on the resulting spark trail to determine whether the puck maintained constant velocity. Before examining the claims the authors make, it is important to note that the interviews did not follow a strict protocol, and the transcript excerpts provided in the article indicate that interviewer intervention may have had considerable impact on the statements made by students.
Several themes emerge from Coelho and Séré’s analysis of the transcripts. Most students express belief in the existence of a ‘true value’ of a measurement. Some of these students express the opinion that measurement of the ‘true value’ is possible and are troubled by the existence of variability in their data. The researchers suggest that student belief in a true value may act simultaneously as an aid and as an impediment to a scientific view of measurement:

The outcome of our survey has demonstrated that the belief in a ‘true value’ is rather fruitful. It encourages the pupils, when measuring, to improve the result through, for instance, the development of original procedures and the repetition of measurements, trying to obtain one, and only one result, thus eliminating variability. On the other hand, the wish to ‘eliminate’ variability implies an incorrect idea about reality, that is, the possibility of obtaining a measurement without any inaccuracy, with the use of ideal instruments and advanced methods, or with the use of measurements provided by extremely well trained scientists. (Coelho and Séré, p.93)

The researchers also note that student perceptions and beliefs play a role in student behavior. They suggest that if the answer to the problem is ‘known’ to the student, the student may use analysis techniques which shortcuts effort. For example, some groups interviewed make a conclusion without making any measurements! Alternatively, some interpret the measurements in ways that confirm held beliefs about the outcome of the experiment. This phenomenon has been observed elsewhere (Schauble 1996).

2.2.2 Lubben et al.: Multiple measurements and models of student thinking

In a series of three studies (Lubben and Millar 1996; Allie, Buffler et al. 1998; Lubben, Campbell et al. 2001), Fred Lubben and various collaborators examine students’ reasoning about the reliability of experimental data. The structures of the three studies are virtually identical. Students respond to a series of six to nine questions concerning repeated measurements. The responses are then analyzed to identify typical patterns of student reasoning. The products of this work are a simple model for the progression of student reasoning concerning the reliability of experimental data and a set of written test items that can be used to assess student reasoning concerning repeat measurements.
In 1996, Lubben and Millar (Lubben and Millar 1996) report the results of a cross-sectional study of about 1000 British School children. Students from three different age groups (roughly equivalent in age to 7th, 9th and 11th graders) responded to a written test comprised of six questions concerning the reliability of experimental data. The four questions for which analysis is reported cover the following topics:

- reasons behind performing repeated measurements,
- identification and treatment of outliers,
- meaning of spread in a data set

Each question presents a discussion between two or three cartoon characters concerning how to proceed during an experiment. The student is asked which character they agree with most closely and why. Student responses to each question are coded according to common themes for that question and the response patterns from the three age groups are compared.

Lubben and Millar’s results show a general movement towards more expert-like understanding with age. In particular, the appreciation of multiple measurements as a way to accommodate scatter in data and construct a mean value increases dramatically from 7th grade (about 25% of the respondents) to 11th grade (about 70% of respondents). The proportion of students who identify an outlier from a list of repeated measurements and choose to remove it from calculations also increases somewhat (from 10% to 20%). The most interesting result from this study concerns a question on the meaning of spread. Two lists of data with the same mean but different spreads are presented and the student is asked to choose which set of data is more reliable. The proportion of students who select the list of data with the smaller spread does increase with age (from 15% to 39%). However, the proportion of students who say that the two sets of data are equally reliable remains stable and large (about 50% of all respondents) for all age groups. The vast majority cites equality of the means as the basis for their decision. Based on the overall results, Lubben and Millar propose an eight level model for the progression of student understanding for empirical evidence. This model will not be described here because it has been revised in later studies.

In 1998, Allie, Buffler, Kaunda, Campbell and Lubben (Allie, Buffler et al. 1998) report a follow-up study conducted at the University of Cape Town, South Africa. Again, the researchers administered a set of questions about multiple measurements. The format and focus of the questions is very similar to the
questions from the 1996 study. However, several changes were made. All six questions described in the article deal with a single experiment, shown in Figure 2-1.

**Figure 2-1. Ball and ramp context for questions about multiple measurements.**
The questions about repeated measurements in Allie et al. (1998) and Lubben et al. (2001) all refer to the experimental apparatus in the picture above. The figure is reproduced from Lubben et al. (2001).
A question concerning the agreement of two sets of experimental data was added. The sample of students was 121 1st year university students enrolled in the Science Foundation Program (SFP), which is designed to redress racial inequalities in educational opportunity under South Africa’s apartheid government. A large proportion of these students spoke English as a second language and had limited lab experience. A coding scheme for student responses is described and inter-rater reliability is reported at 95% between two raters on a sample of 20 randomly selected papers.

The results of the 1998 study largely follow the results of Lubben and Millar’s 1996 study. The South African university students demonstrated a higher level of understanding, with most of the SFP students falling in the top three levels of the eight level model proposed in the cross-sectional study. Most of the students propose taking multiple measurements for time measurements (64% of respondents) and for distance measurements (50%) to construct a mean, but very few (10%) suggest that the purpose of multiple measurements is to determine the spread of the values. Similarly, very few students (28%) mentioned the role of spread when comparing two data sets. Due to the addition of the question on comparing the means of two data sets, the researchers add a level to the model of progression proposed in the 1996 article, but it is clear that not many of these college students appreciate the role that spread plays in the analysis of repeated measurements.

An almost identical study is reported by Lubben, Campbell, Buffler and Allie in 2001 (Lubben, Campbell et al. 2001). First 1st year students (N=257) from the University of Cape Town’s SFP program respond to seven probes, five of which are identical to those described in the 1998 article. The two new questions test ask students to

- report a single number to represent a list of repeated measurements and,
- draw a straight line students to fit a given graph of experimental data.

The new questions are designed to see how students process data. The authors abandon their previous categorization scheme for student responses. Instead, they propose a new, simple three level scheme to categorize student reasoning. According to this scheme,

- **Point reasoning** is characterized by the belief that any measurement could possibly be the exact true value of the measurand.
• **Basic set reasoning**, by contrast, is characterized by the belief that each datum represents only an approximation of the true value.

• **Deep set reasoning** is an extension of basic set reasoning. Students exhibiting deep set reasoning apply the concept of spread to compare data sets and assess the reliability of experimental data.

Responses to five of the seven questions are used to distinguish point reasoners from set reasoners. The two questions where set reasoning is implicit in the problem statement are used to distinguish basic set reasoners from deep set reasoners.

The results in the 2001 article corroborate the earlier findings of Lubben et al. While a large fraction of students will repeat measurements to find an average, understanding of the purpose of determining the spread in data is rare among the students. The data show that more than 60% of the students failed to consider the spread at all when comparing the means of two sets of repeated measurements and fewer than 2% of the students demonstrated deep set comparison on both data comparison probes. The authors separate the questions according to three contexts: collecting, processing, and comparing experimental data. Student reasoning, as determined by the three tier scheme, is reasonably consistent across the contexts of collecting, processing and comparing data, providing support for the new model of student thinking. The researchers, however, also note instances where student behavior is inconsistent with student reasoning. A student may suggest taking multiple measurements to find an average, but then select the mode from a list of repeated measurements.

2.2.3 **Deardorff: Treatment of uncertainty by students**

In a doctoral dissertation, Deardorff (Deardorff 2001) documents common practices of university undergraduate students and graduate students using a variety of instruments covering a myriad of topics related to measurement and measurement uncertainty. Rather than focus students’ notions about the reliability of data like Séré and Lubben, Deardorff examines the difficulties university students (and some experts) have making measurements and using dealing with measurement uncertainty. The data analyzed in the study are principally the written responses of students to lab practical examinations and surveys. Analysis of student lab reports provides additional information about student practices. The subjects are mainly students from two
American large public universities: North Carolina State University (NCSU) and the University of North Carolina at Chapel Hill (UNC).

Deardorff’s conclusions concerning student understanding of uncertainty generally support the findings of the previous research. Many students do not understand the purpose for estimating the inherent uncertainty in empirical data or results. Deardorff bases this conclusion on two observations:

- Students are reticent to produce numerical uncertainty estimates for direct measurements and calculated quantities, *even when explicitly asked to do so on an exam.*
- Students often judge the agreement or nonagreement of empirical results without any reference to the inherent uncertainty of the values being compared.

These difficulties exist among American college students even after one or two semesters of detailed instruction on how to estimate uncertainty and how to use uncertainty to compare results.

Deardorff notes an interesting difference between how students treat directly measured quantities and how they treat calculated results. The values that students give for directly measured quantities (e.g. the diameter of a penny) are generally accurate and the uncertainty estimates, when reported, are generally reasonable. On items involving direct measurement, a vast majority of students uses the correct number of significant figures to express the value. Those who report uncertainty estimates for direct measurements also use no more than two significant figures for the uncertainty.

However, students have considerable difficulty with tasks involving calculations based on measurements. The results of tasks that require calculations from measurement are often inaccurate and the uncertainty estimates are often unreasonable. This result is not surprising. Practical tasks that involve calculation typically require considerable knowledge of relevant underlying science. An example from Deardorff’s study illustrates the point. Students were asked to find the acceleration due to gravity, g, from the motion of a pendulum. Knowledge of the operational definitions of pendulum length and period implicit in the formula used to calculate g are required to make the appropriate measurements. Additionally, conceptual understanding of the pendulum aids in making choices about measurement procedures designed to reduce uncertainty. For instance, the common practice of timing several swings rests on understanding that the period of the pendulum does not change as the pendulum’s motion dies out. Student performance on content-rich
measurement tasks is hard to interpret, because it depends on the interaction of knowledge from the physics and measurement domains.

More interesting is Deardorff’s observation that students often use too many significant figures to report calculated quantities (and their uncertainties). Even students who report uncertainty estimates often quote the uncertainty to three or four significant figures. This is puzzling since most students use significant figures properly when reporting direct measurements and their uncertainties. The reason for this is not clear. Deardorff speculates that this finding might indicate students view the result of a calculation as less inherently uncertain than the measurements on which the calculation is based.

As part of the study, Deardorff compares groups of students and TA’s from two different American universities. The results are highly suggestive. Lab instruction at UNC places considerable emphasis on uncertainty analysis while lab instruction at NCSU largely ignores this aspect of experimentation. Not surprisingly, UNC students report uncertainties in both measured and calculated quantities at much higher rates than their NCSU counterparts, and, in some cases, at higher rates than NCSU TA’s. On some items, UNC students make more accurate measurements than the NCSU students did. However, students at NCSU remove anomalous results when calculating an average more frequently than UNC students. However, these differences cannot be attributed directly to differences in instruction because no pretest was given. It is possible that the observed effect was due to differences in student population at the two institutions.

2.2.4 Other efforts

The literature search revealed two efforts outside the traditional literature pertaining to student understanding of measurement. Soh, Fairbrother and Park (Soh, Bob Fairbrother et al. 1998) study Korean high students’ ability to make accurate distance, time, volume, and force measurements. Soh et al. report that students are not confident in the measurements they make. The Korean students tend to ascribe this lack of confidence in the measurement to ability of the measurer, rather than the features of the measurement tool or task. When asked if a perfect measurement is possible, twelve of 29 students answer affirmatively. Of these, seven cite reasons that focus on the role of machines in measurement. Of the 29 students, only one acknowledges the fallibility of machines.
Evangelinos, Psillos, and Valisseades (Evangelinos, Valissiades et al. 1998) report the results from three of six questions on a survey administered anonymously to 32 first year physics students at the University of Thessaloniki before any lab instruction. About 85% of these students reject the notion of an exact measurement for reading both analog and digital instruments. In justifying their stated opinions, students cite the notion of “approximate quantity” frequently. In another item on the survey, the students are asked to select one of the two positions concerning measurement and then justify their choice. The following translation of the original question is presented in the article:

a) For the researcher, there are only two possible outcomes: a measurement result may be either right or wrong. There is no intermediate situation.

b) For the researcher, there may be three outcomes: a measurement may be right, wrong or a third outcome. Which is this? (Evangelinos, p.2)

Students are equally divided between the two positions. Students adopting the ‘right/wrong’ perspective provide justifications like “this is the aim of the experiment, to resolve the issue whether a hypothesis is correct or not” and “…a quasi-correct result cannot contribute anything in research.” Students espousing the ‘third outcome’ position often give a pragmatic view, suggesting that only an approximate view of reality is possible. The Greek researchers propose a modification on Lubben’s point/set model for students’ notions about the nature of measurement. They suggest that a student’s perception of measurement evolves from a ‘point’ perspective, where measurements are exact, to an ‘approximate’ perspective, where approximations are necessary to come closer to the ‘true value,’ into an ‘interval’ perspective, where the outcome of a measurement lies within an interval. The authors suggest that students may adopt “approximate determinism” to avoid issues associated with probabilistic interpretation of empirical evidence.

2.2.5 Synthesis of previous research

Over the last 10 years, researchers have begun to examine the students’ notions and practices concerning measurement and uncertainty. The general aim of the studies done so far is to document student difficulties associated with measurement and uncertainty and to identify the thinking patterns that underlie student practices. The studies employ a variety of methods, ranging from written questionnaires and tests to observation of student practices during lab activity. Much attention has been paid to students’ perceptions and practices concerning repeated measurements. The studies have been done in different countries with students
of various ages and levels of experimental experience. Taken together, the body of research indicates that high school students, many university undergraduates, and even some graduate students possess notions about measurement that are substantively different from those of experts.

From the cornucopia of student notions and practices documented by the research, a picture of student understanding is beginning to emerge. Recently, a general model for student understanding of measurement and uncertainty was proposed. The model suggests that students fall into one of three categories, based on how they view the nature of a measurement. Students at the lowest level of the model view the result of a measurement as point-like and exact. For these students, any measurement could yield, at least in principle, the “correct answer.” Students in the intermediate level recognize that determining the “correct answer” is impossible. These students apply techniques to produce a best approximation for the “correct answer,” but fail to evaluate the uncertainty of the measurement. Students in the top level adopt an interval view of the measurement. These students recognize that measurements are inherently uncertain and that the task of measurement involves estimation of both the true value of the measurand as well as the uncertainty.

This general model is intriguing for several reasons. This model subsumes a similar model proposed for student views concerning multiple measurements. Many of the myriad difficulties reported in the rest of the literature can be interpreted using this plausible model. If valid, this model could serve as a guide for the development of effective curricular materials. However, the validity of this general model has not been established. If a student’s view of measurement is the underlying cause of misunderstandings, then the student should display similar symptoms in various contexts. With the exception of Deardorff, each study found in the literature focuses on a single aspect of measurement: Séré at al. on measurements taken in the context of a specific kind of lab task, Lubben et al. on students’ notions about multiple measurements as measured by paper and pencil questions, Dimitris et al. on reading a ruler. A study that examines how students’ performance correlates across a variety of contexts is needed to provide support for a global model for student understanding of measurement.

The research done so far provides some information about the impact of lab instruction. The research indicates that student understanding increases with age and relevant experience. For instance, Lubben and Millar’s cross-sectional data show students in higher grades have fewer difficulties understanding the purpose
and treatment of multiple measurements. Similarly, Deardorff’s study shows that graduate TA’s outperformed their undergraduate counterparts. The research also shows that many students have significant misunderstandings even after explicit instruction. The bottom line is that current teaching practice appears to have some effect on student understanding. The question is one of extent.

The present study adds to the existing literature in two fundamental ways. This study examines how student reasoning correlates across several contexts, so that an emerging model for student understanding of variability can be evaluated. Previous studies, primarily aimed to document student difficulties, did not compare patterns of student reasoning across different tasks. Unlike previous studies, the present study incorporates pretest/posttest design. The studies described in this literature review only provide snapshots of student understanding. None of the studies directly measures student learning. A study that incorporates pretest/posttest design is necessary to evaluate the impact of instruction in a convincing way. Further, a study with pretest/posttest design that compares different methods of instruction may identify the salient features of successful teaching.
Chapter 3: Methodology

A test for assessing student understanding of measurement and uncertainty has been developed and administered. The development of the test objectives and items was guided by previous studies on student understanding of these topics. In a pilot study, a preliminary version of the test was administered to about 20 students in a clinical setting. Students were asked to think aloud as they worked through the test. Based on these interviews, the test was revised. The bulk of the data comes from large-scale administration of the revised test to students in classroom settings. About 400 students from two American land grant universities enrolled in various first semester introductory physics lab courses took the test before and after one semester of lab instruction. The test results were analyzed using a mixture of qualitative and quantitative methods. The pretest results were examined to determine what patterns exist among student responses. The pretest and posttest responses were then compared. The student responses were used to evaluate the reliability of the test.

3.1 Sample Description

Students from several populations participated in the study during the Fall 2001 and Spring 2002 semesters. These student groups are described in this section.

All of the participants in this study are students at one of two large public research universities: North Carolina State University (NCSU) and University of North Carolina at Chapel Hill (UNC). The SAT scores and high school rank data for the entire 2001 freshman class (Dept. 2002; Research 2002) show that UNC freshmen have somewhat better academic records than NCSU students.
Table 3-1. SAT math and verbal scores at UNC and NCSU.
Each table entry is the percent of entire 2001 class of first time freshmen with scores in each range on the verbal and math sections of the SAT.

<table>
<thead>
<tr>
<th>Score</th>
<th>SAT Verbal UNC</th>
<th>SAT Verbal NCSU</th>
<th>SAT Math UNC</th>
<th>SAT Math NCSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>700-800</td>
<td>18.2</td>
<td>6</td>
<td>21.5</td>
<td>12</td>
</tr>
<tr>
<td>600-699</td>
<td>46.7</td>
<td>31</td>
<td>48.9</td>
<td>41</td>
</tr>
<tr>
<td>500-599</td>
<td>29.8</td>
<td>48</td>
<td>25.4</td>
<td>38</td>
</tr>
<tr>
<td>400-499</td>
<td>5.1</td>
<td>14</td>
<td>4.1</td>
<td>8</td>
</tr>
<tr>
<td>300-399</td>
<td>0.2</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0-299</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3-2. Composite SAT scores at UNC and NCSU.
Each table entry is the mean SAT score for the entire freshman class at each university. Mean SAT subscores for UNC were estimated from Table 3-1 by taking the midrange score multiplied by the fraction of students with scores in the range. NCSU’s average subscores are published in NCSU’s Common Data Set.

<table>
<thead>
<tr>
<th></th>
<th>UNC</th>
<th>NCSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT Verbal</td>
<td>628</td>
<td>573</td>
</tr>
<tr>
<td>SAT Math</td>
<td>638</td>
<td>602</td>
</tr>
</tbody>
</table>

Table 3-3. High school rank of incoming freshmen at UNC and NCSU.
Each table entry shows the fraction of all 2001 first time freshmen with high school class rank in each category.

<table>
<thead>
<tr>
<th>HS Class Rank</th>
<th>UNC</th>
<th>NCSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top tenth</td>
<td>64 %</td>
<td>35 %</td>
</tr>
<tr>
<td>Top quarter</td>
<td>90 %</td>
<td>74 %</td>
</tr>
<tr>
<td>Top half</td>
<td>98 %</td>
<td>97 %</td>
</tr>
</tbody>
</table>

Enrollment data (Dept. 2002; Research 2002) shown in Table 3-4 points out gender differences in the undergraduate student body.

Table 3-4. Gender balance at UNC and NCSU.
Table entries show numbers (percentages) of total undergraduates by gender.

<table>
<thead>
<tr>
<th></th>
<th>UNC</th>
<th>NCSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>6,128 (40%)</td>
<td>13,029 (58%)</td>
</tr>
<tr>
<td>Female</td>
<td>9,285 (60%)</td>
<td>9,389 (42%)</td>
</tr>
<tr>
<td>Total</td>
<td>15,413</td>
<td>22,418</td>
</tr>
</tbody>
</table>

At UNC, undergrad women outnumber their male counterparts by a ratio of 3:2. At NCSU, which has a somewhat larger undergraduate enrollment, the reverse situation exists.
Six samples of students from UNC and NCSU are listed and described below. Students in sample 1 were enrolled in the second semester of the two semester, calculus based, introductory physics course (PY208) at NCSU in Fall 2002. These students were paid volunteers who took the test in a clinical setting. All students in samples 2 through 6 were enrolled in a first semester introductory mechanics course in Spring 2002, either at UNC or NCSU. Students in samples 2-6 took both the pretest and the posttest during the regular lab period. At UNC, the posttest was given as part of a graded lab practicum. The lab practicum counted as a single lab (10% of the total lab course grade). At NCSU, the posttest was given in stand-alone form. All NCSU students received some class credit for taking the post-test, but the grade incentive varied from sample to sample as described below.

**Sample 1: Think-Aloud Interviews (N=20)**

Subjects for the think-aloud interviews were paid volunteers recruited from two experimental (SCALE-UP) sections of second semester, calculus-based introductory physics (PY208) at North Carolina State University in December 2001. Some of the students had been enrolled in SCALE-UP (Beichner 2002) sections for the first semester course had previously been enrolled in SCALE-UP PY205. Except for the students who had taken SCALE-UP PY205, these students had little or no formal instruction on uncertainty analysis.

**Sample 2: UNC P26 students (N~60)**

These students were enrolled in the calculus-based P26 at the University of North Carolina. All sections of P26 are taught in the traditional lecture/lab format. Students enrolled in the spring semester are primarily on-sequence freshmen. Although 119 P26 students took the pretest, there are only 64 P26 students whose posttest scores are included in the study. This is not due to attrition. One of the P26 TA’s failed to return student papers to the UNC lab director despite repeated entreaties to do so.
Sample 3: UNC P24 students (N~160)

Students enrolled in the algebra-based P24 at UNC are typically life science majors and social science majors. All sections of P24 are taught in the traditional lecture/lab format. There were 191 P24 students who took the pretest and 158 P24 students who took both tests. This loss is mainly due to course attrition.

Sample 4: NCSU PY205T (N=21)

These students were enrolled in traditional, lecture-based sections of PY205 at North Carolina State University in Spring 2001. Students taking the calculus-based PY205 in the spring are typically on-sequence freshman engineering majors. Students from Sample 4 who took the post-test got extra credit worth about 8% of the lab course grade. There was only one section of PY205 lab included in the study. Of the 22 students in the section, all but one took both tests.

Sample 5: NCSU S-UP (N~120)

These students were enrolled in one of two experimental, activity-based (SCALE-UP) section of the calculus-based PY205 at NCSU in Spring 2001. This is the nominally the same course as described in Sample 4, but taught with different techniques. The course has a strong focus on group work, active learning and computers. Like sample 4, sample 5 consists predominantly of on-sequence freshman engineers at NCSU. Some students in this sample may have chosen to enroll in these experimental sections based on the recommendations of friends. A faculty member who often teaches this class is an extraordinarily popular and effective teacher. Students got extra credit for taking the test. Each student in Sample 5 could earn up to about 8% of the total lab course grade from the post-test. The amount of credit, however, depended on the student’s performance on the test. Of the 184 students who took the pretest, 127 took the posttest. Some of this loss is due to students who dropped the course; some of the loss is due to the timing of the posttest. Since the posttest was given in the week before finals, some SCALE-UP students failed to take the posttest because they skipped class that day.
Sample 6: NCSU PY211 (N=55)

These students were enrolled in traditional, lecture-based sections of PY211 at North Carolina State University in Spring 2001. Students taking the algebra-based PY211 are typically life science and social science majors. Students from Sample 6 who took the post-test got extra credit worth about 8% of the lab course grade. Almost all (54 of 58) PY211 students who took the pretest also took the posttest.

3.2 Description of Lab Courses at UNC and NCSU

All the students who took the test during the large-scale administration of the test were enrolled in the first semester introductory physics lab. Students within each sample took the same lab course. This section will describe the lab activities conducted by students in each sample.

UNC P26 (Sample 2)

Students enrolled in P26 at UNC performed nine lab exercises and a lab exam during the semester. The lab exercises covered a wide array of topics in introductory mechanics. There was a strong emphasis on uncertainty and uncertainty analysis. The lab manual (Deardorff 2001) contains a 20 page introduction to measurement and uncertainty. Topics covered in the introduction include estimation of uncertainty (for single and repeated measurements), propagation of uncertainty, random and systematic error, fractional uncertainty and significant figures, and the criterion for agreement of two measurements. Students performed uncertainty calculations in every lab exercise. The first lab exercise the P26 students did (entitled “Is the water safe to drink?”) focused primarily on uncertainty analysis. Students measured the time for a ball a fixed distance through two different fluids. Students conducted multiple trials, calculated standard deviations and evaluated the agreement between two experimental results. Questions about uncertainty analysis and measurement appeared on the lab exam.

UNC P24 (Sample 3)

The lab experience for PY24 students was very similar to the program for P26 students. Students performed eight labs and a lab exam. The emphasis on uncertainty analysis was heavy. Students performed
uncertainty calculations in every lab. The P24 students performed many of the same labs the P26 did. However, P24 students did not do a lab specifically focused on uncertainty.

**NCSU PY205T (Sample 4)**

Students enrolled in a traditional section lab for PY205 performed six lab exercises during the semester. Students followed a prescribed procedure, completing data tables given in the lab manual. Little attention was paid to uncertainty analysis in the traditional PY205 labs. While the lab manual includes instructions on uncertainty estimation and propagation as well as instruction on significant figures, students were rarely required to estimate the uncertainty in measured or calculated quantities. In many of the labs, students calculated the percent difference between a “theoretical” value and an “experimental” one. Students were not required to evaluate the agreement of the two values using uncertainty estimates.

**NCSU PY205S-UP (Sample 5)**

Students enrolled in PY205S-UP did six lab exercises during the semester. Students were given considerable freedom concerning procedures and analysis methods. Rather than providing a recipe for procedures and analysis methods, the instructions for each lab presented students with a goal for the lab. Explicit instruction on estimating uncertainty was given for stacking, repeated measurements and propagation of uncertainty. Detailed information about uncertainty analysis was provided on the class website. Students in the SCALE-UP section of PY205 were responsible for learning uncertainty analysis. The grading rubric for the lab reports included points for uncertainty analysis. It should be noted that each group submitted a single report for each lab exercise, so not every individual was accountable for the contents of the report. To insure individual accountability, questions about uncertainty analysis appeared on class exams, individual homework assignments, and quizzes. A treatment bias exists for this sample. I taught one of the two sections of PY205S-UP and my graduate advisor, Bob Beichner, taught the other. I developed or refined much of the lab curriculum in SCALE-UP PY205 over the period several semesters, up to and including Spring 2001.
NCSU PY211 (Sample 6)

The lab course for students enrolled in PY211 was almost identical to the course for students enrolled in PY205T (Sample 4). Both courses used the same manual (Egler and Mowat 2000) and performed six labs. Five of these labs were the same for both groups. The only difference is that PY211 students did a lab on rotational equilibrium instead of the lab on rotational dynamics that the PY205T did.

3.3 Test Construction

The test developed for this study contains a mixture of multiple choice and free response items covering basic measurement topics. It is designed to elicit student notions about measurement and variability and assess their performance of simple measurement tasks. The instrument contains practical measurement tasks as well as paper and pencil exercises. Several items used in the instrument are adapted from published sources (Allie, Buffler et al. 1998; Evangelinos, Valissiades et al. 1998; Deardorff 2001). The rest were created for this study, based on experience and research on student understanding of measurement. The length of the test was limited due to teaching time constraints, so the test was designed to take 30 to 45 minutes to complete.

3.3.1 Focus of the test

The test focuses on student notions of variability and direct measurements. The test does not include items that address more formal measurement topics (such as inferential statistics and the propagation of uncertainty). At first glance, such a test might seem too easy for college students. Research (Allie, Buffler et al. 1998; Evangelinos, Valissiades et al. 1998; Lubben, Campbell et al. 2001), however, indicates that simple questions about direct measurements are sufficient to elicit a wide array of student beliefs about the inherent variability of measurement that are substantially different than those of experts. Such questions have been successfully used to distinguish between students of different ability. Further, it appears that students’ notions about the inherent variability of measurements underlie almost all of the observed student difficulties. More advanced measurement tasks often involve the interplay between students’ understanding of measurement and students’ understanding of relevant physics topics. For this reason, performance on such tasks may not accurately reflect the student’s understanding of measurement. Another indication that questions on more
advanced topics may not accurately reflect student understanding of measurement concepts comes from a more general finding from physics education research: students can perform well on traditional physics tasks without understanding the underlying concepts. The most conspicuous example is Eric Mazur’s experience (Mazur 1997) with Harvard students who excelled on physics tests with complicated problems but did poorly on a simple multiple choice concept test. A test that focuses on basic concepts avoids these problems.

3.3.2 Repeated measurements, reading scales and stacking

Three contexts for eliciting student thinking about the inherent variability of measurements were chosen for study. The extensive work of Lubben and others clearly shows that student responses to questions about repeated measurements give valuable insight into students’ notions about variability. Although the evidence is less clear, the work of Evangelinos et al. suggests that questions about reading a ruler may provide similar information. The third context that might provide valuable insight into students’ notions about variability had not yet been studied.

Consider the task of estimating the thickness of a page of this dissertation using a ruler. Measuring the thickness of, say, 100 pages together and dividing that measurement by 100 provides a far less uncertain value for the thickness of a typical page than simply trying to measure the thickness of the page directly. Only one (Taylor) of three popular error analysis texts (Bevington and Robinson 1992; Baird 1995; Taylor 1997) mentions this practice, yet the practice is often used in elementary labs with time, distance and mass measurements. For instance, students are often instructed to use stacking when measuring the period of oscillation for a simple pendulum. Since the technique is unnamed, I will coin the word stacking, and use the word to describe the technique of using the total measurement of N similar things all at once (placed “end-to-end” in some sense) divided by N to represent the value of the measurand. Stacking effectively decreases the uncertainty in the average due to least count of a measurement device by “spreading out” the absolute uncertainty over the entire sample. If the variability of the entity being stacked is small compared to the least count uncertainty, the average from stacking is a far better estimate for the measurand than the value obtained from a single measurement of a single object. Notions about the nature of measurement and variability clearly play a key role in stacking, but until this study educational researchers had not yet examined student thinking about stacking.
3.3.3 Item details

During the development of this test, several items were revised, mainly due to concerns raised during the pilot study. The questions presented in this section represent the test items as they appeared on the large-scale administration of the test. The items have been modified slightly to improve the readability of the dissertation: (1) Spaces for student explanations have been reduced or eliminated and (2) titles for each question have been added for reference. The test that the students took is shown in Appendix A.

The first five questions were adapted from the 1998 study by Allie et al. Several modifications were made. In the South African study, three cartoon characters discussed how to proceed at various points during the experiment involving a ball and ramp previously shown in Figure 2-1. Because the NCSU students enrolled in the traditional lab course often perform the ball and ramp experiment during the semester, the ball and ramp experiment was replaced by the “dropping balls in the sand” experiment shown in Figure 3-1.

An experiment is performed by students in a physics laboratory. A glass marble is dropped into a container of sand. The ball is released from a height \( h \) above the surface of the sand. The ball is carefully removed from the sand, leaving a crater of diameter \( D \). The students have been asked to investigate how the diameter of the crater changes when the height \( h \) is changed. A meter stick is used to measure \( D \) and \( h \). The students work in groups on the experiment.

![Figure 3-1. Marble and sand context for questions about multiple measurements.](image)
The questions about repeated measurements in this study involve the apparatus above. The text above the diagram is the description given on the test paper. The apparatus and experiment are taken from Amato and Williams (Amato and Williams 1998).

The present study used five of the six questions described in the report of the South African study. A question about repeating time measurements was dropped, because the results from the South African study show that this question does not provide much insight into students’ notions of variability.

The language of the questions was changed as little as possible from the South African study, but some changes were made to the presentation. The cartoon figures were removed and replaced with letters (see examples below). In the South African study, each question was presented on a single card. Students turned in one card before receiving the next question to insure that students did not do back to change their answers to previous questions. This was deemed impractical given the scale of the present study. Instead, questions were presented in order on paper and students were asked to work these questions in order. During the large-scale administration of the test, I carefully observed five sections of students taking the test. I did not see any students reading ahead or doing the first five questions out of order. An examination of the test papers provides no evidence that students changed early answers based on responses to later questions.

The questions adapted from Allie et al. are presented below. On the test, each question occupies at least half a page to provide space for student explanations, since reasons that students provide to support their letter choice are the primary data for these questions. Questions 1 through 5 all refer to the experimental apparatus in Figure 3-1. Question 1 and question 2 deal with the reasoning behind multiple measurements.

---

**Question #1: Repeat distance measurement (RD)**

1. The students have to determine \( D \) when \( h = 2500 \) mm. One group releases the ball at a height \( h = 2500 \) mm and, using a meter stick, they measure \( D \) to be 124 mm. The following discussion takes place between the students.

   **A:** I think we should drop the ball a few more times from the same height and measure \( D \) each time.
   **B:** Why? We’ve got the result already. We do not need to do any more drops.
   **C:** We must drop the ball just one more time from the same height.

   With whom do you most closely agree? (Circle one):  A  B  C

   Explain your choice.
Question #2: Repeat distance measurement again (RDA)

2. The group of students decides to release the ball again from \( h = 2500 \) mm. This time they measure \( D = 132 \) mm.

First release: \( h = 2500 \) mm \( D = 124 \) mm  
Second release: \( h = 2500 \) mm \( D = 132 \) mm

The following discussion takes place between the students.

A: *We have enough. We don’t need to repeat the measurement again.*  
B: *We need to release the ball just one more time.*  
C: *Three releases are not enough. We must release the ball several more times.*

With whom do you most closely agree? (Circle one): A B C

Explain your choice.

While Allie’s study shows that many students indicate that repeat measurements are needed, the reasons they give for the practice are not uniform or expert-like. Generally, experts view the purpose of multiple measurements as two-fold: to get a best value and to estimate the uncertainty in that best value due to variability. As more data is taken, estimates of a best value and the uncertainty become increasingly reliable. Experts also recognize that multiple measurements using the same procedure are insufficient to detect or estimate uncertainty arising from systematic sources.

Questions 3 through 5 deal with interpreting the data from repeated measurements. Question 3 specifically addresses whether to include an outlier when calculating a mean:

Question #3: Treatment of outlier (AN)

3. A group of students has to calculate the average of their measurements after taking six diameter readings: Their results are as follows (mm): 128, 123, 115, 172, 119, 132. The students discuss what to write down for the average of the readings.

A: *All we need to do is add our measurements and then divide by six.*  
B: *No. We should ignore 172 mm, then add the rest and divide by five.*

With whom do you most closely agree? (Circle one): A B

Explain your choice.
Experts use a variety of tests to identify outliers. Despite the variety of tests, experts universally remove outliers when calculating a mean, because inclusion of the outlier produces a skewed value for the mean. The distribution for the data presented in the problem is shown below:

![Question #3](image)

**Figure 3-2. Visualization of the data presented in question #3.**

The utility of visual representations for qualitative uncertainty analysis has been noted by researchers, but the same researchers note that students do not spontaneously use such representations.

The value 172 mm was chosen for this item because it lies quite far (more than seven standard deviations) from the mean.

Question 4 addresses the implications of the spread (dispersion) of multiple measurements:

**Question #4: Same mean, different spread (SMDS)**

4. Two groups of students compare their results for the diameter measurement.

Group A: 126, 130, 122, 127, 120  Average = 125 mm
Group B: 121, 112, 142, 115, 135  Average = 125 mm

**A:** Our results are better than yours. They are all between 120 mm and 130 mm. Yours are spread between 112 mm and 142 mm.

**B:** Our results are just as good as yours. Our average is the same as yours. We both get 125 mm for the diameter.

With whom do you most closely agree? (Circle one):  A  B

Explain your choice.
Experts universally regard dispersion as measure of the (un)reliability of data. The equality of the two averages in this example is only important because it rules out the possibility that one of these values is closer to a true value of the measurand than the other.

Question 5 regards the condition for agreement of two measurements:

---

**Question #5: Different mean, same spread (DMSS)**

5. Two groups of students compare their results for five releases of the ball for five releases at \( h = 2500 \text{ mm} \).

Group A: 118 125 120 128 124   Average = 123 mm  
Group B: 121 127 122 124 131   Average = 125 mm

A: *Our result agrees with yours.*  
B: *No, your result does not agree with ours.*

With whom do you most closely agree? (Circle one):  
A    B

Explain your choice.

---

When evaluating agreement of two measurements, both the values and the uncertainties must be considered. These two data sets were chosen to have significant overlap, as shown in Figure 3-3.

*Figure 3-3. Visualization of the data presented in question #5.*  
The two samples of data are not significantly different.

The overlap of these confidence intervals can be supported by results of the Student \( t \)-test for comparing small sample means. Using the Student \( t \)-test, the hypothesis of equal means cannot be rejected at the 95% confidence level.
The rest of the items on the test were constructed for this test. Question #7 asks students to read a ruler and choose among alternatives of increasing resolution. Again, students are prompted for the reason behind their choice.

**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.

Choices A through C represent 100% confidence intervals for the measurement; true value for the measurement clearly lies within these boundaries. Choice D is less conservative (one can argue about the confidence interval represented in this choice). Choice D is most in line with expert opinion as expressed by the ISO Guide, which advocates a realistic (rather than “safe”) evaluation of uncertainty. Choice E is unrealistically restrictive, since a number of sources of uncertainty (such as the width of the ruler’s markings) are on the order of one tenth of the smallest division on the ruler.

Students in the pilot study had considerable difficulty with the wording in early versions of this question. The first draft is shown below:
Question #6: FIRST DRAFT (used in pilot study only)

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

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

Explain your choice.

For this version of the question, students responded that the bug was between 0cm and 2cm, apparently confusing the bug’s position for the bug’s length. Students also were not sure which parts of the bug to include in the measurement. A few minor changes seemed to eliminate the confusion. The ruler’s markings were changed, so that the choices could not possibly correspond to the bug’s position. The word “long” was added to each distractor, clearly indicating that the answer is a length. The guidelines marking the ends of the bug’s body were also substantially darkened.

For question #7, students are asked to determine the average thickness of the floppy disks in the picture and select from alternatives of increasing resolution.
Question #7: Reading a ruler with stacking (RRS)

What method should be used to measure the average thickness of the floppy disks in the picture as accurately and precisely as possible?

Which of the following best describes the average thickness of the floppy disks in the picture?

a) between 0 and 1 cm
b) between 0.3 and 0.4 cm
c) between 0.32 and 0.34 cm
d) between 0.323 and 0.327 cm
e) between 0.3248 and 0.3252 cm

Explain your choice.

Students have a number of choices to make before deciding on an answer:

- How should the ruler be read? (see discussion for question #6)
- Should stacking be employed? multiple measurements of single disks?
- What level of uncertainty does the answer have?

The two prompts for explanations are designed to solicit student thinking about these and other issues.

Stacking should be employed, because the uncertainty due to instrumental resolution is large compared to the
variations in individual disk thickness. A measurement consistent with choice D is achieved when all ten disks are stacked together and the ruler is read to nearest 0.01 cm. During the pilot study, several different versions of this question were tested. The version that appears on the test uses two questions about the methods students use for the measurement. While the two questions and the picture appear to suggest stacking as a possibility, many students in the think aloud interviews did not consider this option.

Question #8 presents students with a list of values for a measurement. Students are asked to estimate the mass of a marble from a list of multiple measurements that includes an outlier. Students are also asked to estimate the uncertainty in the mass.

**Question #8: Recognizing and treating an outlier (AN2)**

8. Several students measure the mass of the same marble on a scale that can be read to the nearest 0.01 g. The results are as follows: 50.91 g, 50.90 g, 50.87 g, 58.03 g, 50.82 g. What value should they report for the mass of the marble? for the uncertainty in the mass?

\[
\text{Mass} = \underline{\phantom{0}} \\
\text{Uncertainty in the mass} = \underline{\phantom{0}}
\]

Explain how you determined your answers.

A correct solution requires that the student recognize and remove the outlier. The mass may be estimated by the mean of the values (excluding the anomaly) and the uncertainty can be estimated by the standard deviation of the values (excluding the anomaly).

Students are explicitly asked to supply an uncertainty estimate for questions #8 through #11. The answer blank was provided to avoid non-reporting of uncertainty on the posttest. In the only study in which students were asked to include uncertainty estimates with measurement, Deardorff (Deardorff 2001) found that many students failed to report the uncertainties. While students were explicitly asked to provide uncertainties, the instructions to do so were buried on the first page of a multiple page document among other general instructions. Context clues (such as the blatant one in the item above) on individual test items were absent. The uncertainty blank on the pre-instruction administration of the test will likely yield little information about student understanding. During the pilot study, all students produced a response for each of
the uncertainty blanks, but these responses were varied and mostly incorrect. In most cases, it is unclear whether these responses are guesses, naïve interpretations of the word “uncertainty,” or results of previous instruction. When questioned, most of the students indicated that their confidence in the answers to the uncertainty questions was low.

Questions 9 through 11 involve the use of equipment to make measurements. The nature of measurement tasks depends somewhat upon the available equipment. Student performance on these items therefore depends somewhat on the available equipment. Rather than provide a standard equipment list available to all students, students only had access to the equipment typically used at their lab site. This choice was based on logistical and design considerations. Student performance is likely to be highest on measurement tasks when the student is familiar with measurement tools. The students will be more facile with their use and more aware of the inherent limitations. A set of standard equipment for the exam would necessarily include unfamiliar equipment and would require a large logistical effort to assemble.

In question #9, students measure the diameter of a quarter. Students are prompted for numerical values and reasoning.

---

**Question #9: Measure quarter diameter (QD)**

9. Find the diameter of a typical quarter as precisely and accurately as you can, using the available equipment. Estimate the uncertainty in your measurement.

\[
\text{Diameter} = \underline{\quad}\]

\[
\text{Uncertainty in the Diameter} = \underline{\quad}\]

Describe how you made the measurement(s).

Explain how you determined the uncertainty.

---

The record of student reasoning is particularly important for the performance of measurement items. On this question, students can perform this measurement with different tools (e.g. vernier caliper, ruler) and different methods (e.g. repeated measurements, stacking). The choice of tool and technique directly affect the uncertainty of the measurement.
In question #10, students measure the mass of a typical paper clip. The low mass of the paper clip (compared to the smallest division on the balances available for this task) strongly suggests that stacking be employed to minimize uncertainty from instrumental resolution.

**Question #10: Measure paper clip mass (PCM)**

**10.** Find the mass of an average #3 Gem brand paper clip as precisely and accurately as you can, using the paper clips provided and the available equipment. Estimate the uncertainty in your measurement.

Mass = __________

Uncertainty in the mass = __________

Describe how you made the measurement(s)

Explain how you determined your uncertainty.

The labs at UNC use digital balances with a least count of 0.05 g. The NCSU labs are equipped with mechanical balances with a smallest division of 0.1 g.

**Question #11: Coffee filter drop time (DT)**

**11.** Find the time it takes a single coffee filter to fall a distance of 1.00 meter as precisely and accurately as you can, using the available equipment. (The filter is released from rest). Estimate the uncertainty in your measurement.

Time = __________

Uncertainty in the time = __________

Describe how you made the measurement(s)

Explain how you determined the uncertainty.
Both schools use the same electronic stopclock in the lab. The device has a least count of 0.01 s. The existence of many random sources of uncertainty strongly indicates the use of multiple trials.

Correct answers for items #9 through #11 do not exist in the typical sense. The somewhat open-ended nature of these items naturally leads to a diverse set of student responses. The quality of the student responses has a variety of components:

- Measurement method employed (stacking, multiple trials, hybrid)
- Accuracy of measurement
- Reasonable uncertainty estimate (for the chosen method)
- Proper use of significant figures

Because these factors are interdependent, analysis of the student responses must necessarily use an inductive grounded theory approach (see Analysis Methods section below for a description of this method). The relative “correctness” of student responses will be analyzed in the Results.

3.4 Think-Aloud Interviews

Before large-scale administration, I conducted think-aloud interviews to see how students interpreted and responded to items on the test as well as to determine estimates for the time students needed for each item. In this type of interview, the interviewee is asked to speak as they perform a task. Participation of the interviewer is limited to minimize threats to validity associated with interactions between the interviewer and interviewee.

A pool of about 25 students was recruited from the two experimental classes described in Sample 1. To recruit students, I circulated a sign-up sheet in each class. Volunteers were paid $8 for an interview about lab skills lasting approximately one hour. A recruitment email was also sent to the class. Interviews were scheduled at the convenience of each volunteer via email during December 2001. Twenty interviews were conducted in all. The interviews, which took place in a private room, were videotaped. The equipment available for the test was the same equipment these students used in lab. Care was taken to restore the equipment to the same state for the beginning of each interview (e.g. zero the scale, place all paper clips in the cup, reset the stopwatch, replace worn coffee filters).
Each interview followed the same format. In the first part of the interview, the basic structure and purpose of the interview were described to the student. This part took less than 5 minutes. Then the following script was followed:

In this interview, I’m going to ask you to think aloud as you answer some questions about measurement. Since you’re probably not used to thinking aloud as you work, I want you to practice thinking aloud on the first two questions. As you work, I want you to say whatever comes to mind. Don’t worry about making sense. Don’t worry about complete sentences. Just say whatever comes to mind. Let’s go ahead and try a practice run.

If the student was silent for a prolonged period (more than about 30 seconds) during this practice task, some coaching on thinking aloud was offered and the second question was used for practice. The coaching typically consisted of encouragement. Students were told to keep talking, not worry about being correct, and then reminded that they could ask questions after the interview. After the practice was complete, the following script was followed:

Great! Now you’re ready for the main task. Work through these questions on measurement. Work through the questions in order. Record any measurements you make on the paper. Remember to keep talking as you work. When you are finished let me know, and we will continue the interview.

If the subject asked the interviewer a question about the work, the interviewer instructed the participant to do their best and reassured them that any questions they had would be answered at the end of the interview. Once the subject indicated that he/she had finished the test, the interviewer asked questions about individual test items. The most common questions asked by the interviewer were:

- How confident are you about your answers?
- Are there any questions that weren’t clear?
- Would you like to change any of your answers? Which? Why?
- I’d like to ask you to explain more about how you did #…
- Are there any questions you want answered about this test?

After the debriefing, the subject was thanked and paid.

Changes were made to the test during the course of the two weeks interviews took place, so not every subject did exactly the same test. Two versions of the questions developed by Allie et al. were used during the interviews. During the first eight interviews, the questions were presented using Allie’s original context (the “ball and ramp” experiment). During the last twelve interviews, the questions were presented using the “dropping balls in the sand” context described above. The interviews uncovered student misunderstanding of items (see discussion of bug and ruler question above) and typographical errors. Test items were revised when
problems were identified or typos were found and the revised test was used for subsequent interviews.

Questions that were modified during the think aloud interviews are Question 6 (Reading a Ruler), Question 7 (Reading a Ruler with Stacking), and Question 8 (Recognizing and Treating an Outlier).

3.5 Analysis Methods

The student responses were analyzed using a mixture of qualitative and quantitative methods. Qualitative analysis of students’ verbal responses were used to gain understanding of the reasoning behind student responses. The verbal and numerical responses were coded into categories. This inductive grounded theory approach was taken with the open-ended responses. A coding scheme based on the observed responses was constructed from themes that emerge from repeated words, phrases or consistent patterns in the expression of numerical data. The coding process was tested for reliability to make sure that different readers categorize responses in the same way. The coding schemes were revised and retested until reliable categorization is achieved. In the case of the questions adapted from Allie et al., the coding schemes employed by the authors of the questions were used as a starting point for the coding scheme.

Quantitative analysis of the data involved inferential statistics. Most of the data is categorical or ordinal (i.e. in ordered categories). Proportions of students falling into a particular category were calculated. Confidence intervals for proportions can be calculated using parametric statistics. For large random samples (see condition below), the standard error of a proportion is estimated by:

$$\hat{\sigma}_p = \sqrt{\hat{p}(1-\hat{p})/n}$$

where

$$\hat{p} = \text{proportion of sample meeting a given criterion}$$

$$n = \text{sample size}$$

$$n > 5 / \max(\hat{p},1-\hat{p})$$

The half width of a confidence interval can be found by multiplying by the value of the z-distribution that corresponds to desired level of confidence. Z-values for commonly used levels of confidence are shown in Table 3-5.
To compare proportions, the hypothesis for equal proportions was used. A z-score is generated as shown below:

$$z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

- $p_i =$ proportion in sample $i$
- $n_i =$ size of sample $i$
- $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

The z-value is then compared to the threshold z-value associated with a predetermined confidence level or $\alpha$-level (see Table 3-5). A difference is deemed “statistically significant” if the probability that two proportions so different could have been sampled from the same population is lower that some arbitrary threshold probability value called the $\alpha$-level. For social science work, an $\alpha$-level of 0.05 is typically used.

The same basic procedure can be used to compare means. The standard deviation for the mean for sufficiently large samples ($n>30$) can be estimated by

$$\hat{\sigma}_\mu = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

where

- $x_i =$ individual data values
\[ \bar{X} = \text{average of the individual data values} \]

\[ n = \text{sample size} \]

Confidence intervals are constructed in the same manner as for proportions. To test the equality of means, a \( t \)-score is calculated according to the following formula:

\[
    t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}
\]

where

\[ \bar{X}_1, \bar{X}_2 = \text{means of groups 1 and 2} \]

\[ n_1, n_2 = \text{sample sizes of groups 1 and 2} \]

\[
    s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}
\]

The \( t \)-score for means is interpreted in a manner similar to the \( z \)-score for proportions. If the \( t \)-value exceeds the threshold \( t \)-value for \( n_1 + n_2 - 2 \) degrees of freedom and a given \( \alpha \)-level, the hypothesis of equal means is rejected.

Much of the analysis involved detecting and quantifying the correlations between one set of data and another (e.g. pretest score and post-test score, response category on question 1 and response to question 5). There are many tests for analyzing correlation, depending on the types of variables (categorical, ordinal, interval) being compared. However, most of these tests use some type of proportional reduction of error scheme. This means that the test statistic is constructed by subtracting the errors made by assuming a relationship between the variables from the errors made assuming no relationship divided by the errors made by the “no relationship” model. The statistical significance of the model is determined by the distribution of the test statistic. In some cases, the test statistic can also be used to calculate some measure of the strength of the relationship. (It is important to remember that even weak relationships can be found statistically significant if large enough random samples are used).
The most famous of this family of correlation measures is Pearson’s $r$, for a presumed linear relationship of two continuous numerical variables (such as pretest score and posttest score):

$$r = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where

$x_i, y_i = x$ and $y$ coordinates for each data point

$\bar{x}, \bar{y} =$ mean values for $x$ and $y$ data

Pearson’s $r$ is directly related to the strength of the presumed relationship, i.e. the slope ($b$) of the line:

$$b = r \sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}}$$

The square of Pearson’s correlation coefficient takes on values from zero (for unrelated variables) to one (for a perfect linear fit) and can be interpreted as the fraction of the data’s variance that is removed by assuming the linear model. If a linear relationship exists, the slope should be significantly different from zero. The test statistic for this condition is

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

where

$n =$ number of data points

$r =$ Pearson’s correlation coefficient

The t-distribution is used to find the width of the confidence interval for the slope of the presumed linear relationship. If the (arbitrarily chosen) confidence interval for the slope does not include zero, the model is considered statistically significant. Before using Pearson’s test, it is important to plot the data to see whether a linear model is appropriate.

Relationships between categorical and ordinal data can also be detected and described. As with interval data, a visual representation of the data should be examined before a statistical test is done. For ordinal and categorical variables, cross-classification tables and mosaic plots are used to display the data.
Consider the following fictitious example. A researcher wants to see if there is a relationship between eye color and handedness among people. A random sample of 120 people is polled, and the results are shown in Table 3-5.

<table>
<thead>
<tr>
<th></th>
<th>Blue-eyed</th>
<th>Brown-eyed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lefty</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Righty</td>
<td>27</td>
<td>81</td>
</tr>
</tbody>
</table>

Table 3-6. Fictitious data for example mosaic plot shown in Figure 3-4.
A fictitious sample of 120 randomly selected people polled concerning their eye color and handedness to determine whether the two traits are related. Each table entry represents the number of people falling into each category.

Notice that the data has been rigged. Notice that 75% of righties and 75% of lefties are brown-eyed. This strongly suggests that the eye color and handedness are unrelated. This is hard to see when the data is presented in tabular form. Mosaic plots, like the one shown in Figure 3-4, are far easier to read.

Figure 3-4. Mosaic plot of the fictitious data shown in Table 3-6.
The height (width) of each rectangle represents the fraction of people who fall in each category along the vertical (horizontal) axis. In the fictitious sample, lefties are equally common among blue-eyed and brown-eyed people. Horizontal lines that cross the entire plot indicate a lack of relationship between the variables.

Each table entry is represented by a different area. The length (height) of each rectangle represents the proportion of the sample in each category along the horizontal (vertical) axis. It is quite easy to see that the proportion of righties does not depend on eye color. For contrast, consider another fictitious example concerning the relationship between car ownership and snowblower ownership in Rochester, NY shown in Figure 3-5.
A fictitious, but randomly selected sample of Rochester, NY residents was polled. The results show that snowblowers are far more common among car owners than people who don’t own cars. Notice that no horizontal line that crosses the entire plot.

In the mosaic plot, it is very easy to see that snowblower ownership is far more frequent among car owners than among people who don’t own cars. Determining whether this relationship is statistically significant requires a statistical test.

Many different tests for a relationship between categorical and ordinal data exist (examples include chi-square, Spearman $R$, Fisher’s exact test and its generalizations, Kendall’s tau). Which one to use depends on the nature of the variables involved (categorical or ordinal) and the number of categories for each variable. All of the tests provide some numerical estimate of the strength of the relationship between the variables, and a test for statistical significance of the model determined by the distribution of some test statistic that depends on the strength of the relationship and the size of the data set. As with Pearson’s correlation coefficient, it is important to keep in mind that statistical significance does not always imply a strong relationship between the variables of interest because small differences can be detected if the random sample is large enough.

Many of the tests for significance listed above are associated with measures that characterize the strength of the relationship. For instance, dividing the chi-square statistic by the number of data points in the sample yields phi-square. This measure of association can have values that range from zero (for no relationship) to one (each variable is a perfect predictor of the other). Phi-square is a “percent reduction in error” measure. This means that phi-square measures the percent decrease in mis-categorization achieved by
using the values in the cross-categorization table rather than assuming the two variables are unrelated. The value of phi-square can be deduced from a mosaic plot without knowing the sample size.

3.6 Reliability and Validity

The primary measurement instrument for this study was a written test. As with any measurement, there are two primary components to the quality of the data generated by the test:

- the degree of consistency with which the test produces the same measurement for the same thing.
- the degree to which the test actually measures what it is supposed to measure.

In the social sciences, these concerns are called **reliability** and **validity**, respectively, and roughly correspond to the ISO concepts of repeatability and accuracy. (The ISO guide deprecates use of the word “precision” and offers no definition for this commonly used term).

Ideally, a social science measure (such as a concept test) should be consistent from day to day, from observer to observer, and from item to item. Once the test has been administered, the reliability can be readily assessed by checking the internal consistency of student responses. Post-hoc measures of reliability abound in the social sciences, often expressed as correlation coefficients. During the coding process, at least two different people categorized responses and the inter-rater reliability was monitored. Once the responses have been coded, other correlations were calculated to assess the reliability of the test. Some examples include test-retest reliability (how well student performance correlates across different administrations of the same test), split-half correlation (how well student performance on half of the test correlates to performance on the other half of the test), and point biserial correlation (how well performance on an individual test item correlates to the overall test score).

Numerical estimates of validity in social science do exist in the form of correlation coefficients between related measures (some examples of correlative measures of validity include predictive reliability, concurrent validity, and construct validity). However, these numerical estimates are hard to produce because there are few validated measures available for comparison. Consequently, correlational measures of validity are rarely used. Instead, the validity is usually “established” by argument. Two types of validity are
particularly germane to this study. Face validity is a subjective judgment of whether the instrument measures what it is supposed to measure. Content validity measures how well the test addresses specific content objectives. Content validity is established by comparing the test items and objectives to external criteria (e.g. textbooks, surveys of teachers, misconceptions research).

3.7 Summary

A test for assessing students’ conceptions about measurement and uncertainty was developed. The test was administered to several different groups taking introductory physics lab courses. The results were analyzed using a mixture of quantitative and qualitative techniques to answer the two basic questions presented in the introduction of this thesis. Validity and reliability of the test was evaluated.
Chapter 4: Analysis and Findings

A test for assessing student understanding of measurement and uncertainty has been developed and administered. In this chapter, the results of the test will be analyzed. In section 4.1 and its subsections pretest and posttest performance of NCSU traditional lab students, UNC lab students and SCALE-UP students are compared. In this early part of the chapter, details of the analysis procedure will be described and the major findings will be presented.

In section 4.2 and its subsections, relationships between different aspects of student performance on the pretest are explored. First, correlations between items that measure the same aspect of performance (e.g. ruler reading, stacking, and outlier exclusion) will be presented to establish predictive validity of each aspect. Correlations between various aspects of performance will then presented. Unless otherwise indicated, only pretests with analyzable responses are included in these analyses. Posttest responses are likely to include the effects of instruction. As a result, analysis of the posttest is unlikely to uncover “naturally occurring” connections between different aspects of student thinking.

Throughout this chapter, minimal description of test items will be provided. For a more complete description of test items, please refer to Chapter 3.

4.1 Comparison of pretest and posttest

In this section, pretest and posttest results are compared. Only students who provide analyzable responses on both pretest and posttest are considered for the comparisons. This leads to maximum sample sizes of 75 NCSU traditional lab students, 127 SCALE-UP students and 224 UNC students. The exact number of matched records varies slightly from question to question, because the number of students who provide analyzable responses on both pretest and posttest varies slightly from question to question. The results for each question are presented in the order the questions appear on the test.

4.1.1 Questions 1 and 2: Purpose of repeat measurements

Questions 1 through 5 on this test were adapted from a published test. In the two research articles where the test items are described, the authors outline categories of student responses. Examples of student responses are given and numbers of student responses falling into each category are reported. These published
categorization schemes were used as a starting point for the analysis presented in this study, but minor changes were made to improve the reliability and validity of the results.

Questions 1 and 2, which concern students’ reasons for taking multiple measurements, ask students to choose whether more data should be taken at two different junctures during a hypothetical experiment. In 1998, Allie et al. (Allie, Buffler et al. 1998) describe six categories of student responses to these questions. A table outlining the categorization scheme from the 1998 article is reproduced verbatim in Table 4-1.

Table 4-1. Student reasoning on questions 1 and 2.
Categorization scheme for student responses to questions concerning the purpose of repeat measurements published by Allie et al.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>No repeats are needed.</td>
</tr>
<tr>
<td>R2</td>
<td>Repeats provide practice to improve the process of taking measurements.</td>
</tr>
<tr>
<td>R3</td>
<td>Repeats are needed to find the recurring measurement</td>
</tr>
<tr>
<td>R4</td>
<td>Repeats are needed to improve the accuracy</td>
</tr>
<tr>
<td>R5</td>
<td>Repeats are needed for establishing a mean</td>
</tr>
<tr>
<td>R6</td>
<td>Repeats are needed for establishing a spread</td>
</tr>
<tr>
<td>R0</td>
<td>Not codeable</td>
</tr>
</tbody>
</table>

In 2001, another article (Lubben, Campbell et al. 2001) published by the same research group implied that the six-category scheme from the 1998 article was used, but results in the 2001 article are reported using a simpler scheme, with only two categories (termed “point” and “set”). According to the scheme, “point” reasoners view any measurement as potentially being the exact true value of the measurand. For “set” reasoners, each datum represents an approximation of the true value.

In general, the responses to questions 1 and 2 given by the students in this study qualitatively match the results published by the developers of the questions. However, initial attempts to classify student responses using the 1998 scheme exposed ambiguity and inconsistency among some of the categories. The main problem concerns students who recognize that taking more measurements improves something about the quality of the data. In the 1998 article, Allie does not clearly articulate whether these students should be in R4 or R6. At one point, the article states:

Category R4 consists of responses which make a very general reference to repeating in order to increase the accuracy. These students write that ‘the larger the number of readings, the greater the accuracy of the times achieved for the experiment’ (RT), and ‘the more measurements you take the
more you know how accurate you are. One or two measurements doesn’t tell you enough about the real time taken’ (RDA).

Later, a similar description for R6 responses is given:

Category R6 comprises those responses suggesting that repeating measurements is needed to improve the spread of the measurements where students suggest that ‘in order to be more precise, that is, reduce the uncertainty, we have to take as many readings as possible’ (RDA). More than half of the responses in this category also mention calculating a mean, for example, ‘For any measurement in physics there will be systematic errors. Hence the value of time in each case will differ. So they will need to find the average time. There will be uncertainty associated with that average of time’ (RT).

In the R6 examples Allie provides, the reader must assume that the student uses the word ‘uncertainty’ to convey its complete technical meaning to support the R6 classification. Given the student confusion concerning terminology reported by Sere (Sere, Journeaux et al. 1993) and Deardorff (Deardorff 2001) (among others), this seems imprudent. In order to ameliorate confusion between R4 and R6 for the present study, only responses that explicitly mention the quantification of the spread of the data are classified as R6. Some examples of R6 responses collected from the present study include:

- It must be dropped a few more times to find an average of D and see how much it changes every time.
- There are many factors that could affect the crater size. Since this experiment is not performed in a vacuum, air resistance comes into play. So dropping the ball a few more times, the more times you could allow to find the margin of error.
- By choosing A, you will be able to record a few different measurements for D. After recording these measurements, you can calculate the uncertainty of D so you can find a more accurate result.

Notice that each of these responses suggests the student is trying to attach a number to the spread: “…how much it changes…,” “…find the margin of error…,” and “calculate [dissertation author’s italics] the uncertainty.” Responses that vaguely remark that more data is “better” in some sense without explicit reference to quantifying the spread of the data are classified as R4. Responses that fit the criteria for both R5 and R6 were coded as R56. All other responses that fit more than one category were put into the highest category consistent with the response.

The categorization of student responses to questions 1 and 2 was a multiple step process. A sample of student responses was first categorized by two coders (DSA and DWC) using the published guidelines. Based on discussions, written guidelines that included descriptions of the categories, examples of student responses,
and notes were developed. Several months later, three graduate students in physics education (DSA, JJM and MAK) coded a small random sample of student responses using the written guidelines. The agreement between the three raters (DSA, JJM, MAK) was assessed and the reasons behind disagreements about the classification of individual student responses were discussed. Within a day of this discussion, the written guidelines were revised and the same three raters categorized a random sample of 100 student responses. This sample of 100 classifications by three raters can be used to assess the inter-rater reliability of the coding. The agreement rate between the three possible pairs of coders ranged from 73% to 82% on questions 1 and 2. More than half of the coding disagreements arose from responses where both raters categorized the student response as reflecting lower level thought (R0 through R4). Less than a day after the training of the coders and the final assessment of inter-rater reliability, the author of this dissertation (DSA) coded the remainder of the responses using the revised written guidelines.

To further improve reliability of the data, the categories were combined into the two categories suggested by the questions’ designers in the 2001 article, as shown in Figure 4-1.

**Figure 4-1. Classifying reasoning as “point” or “set.”**

Categories R5, R6, and R56 are combined and collectively called “set reasoning,” reflecting the view that a set of repeated trials must be collected to construct statistics which describe the measurement. In this study, student responses in categories R0 through R4 are combined and are collectively called “point reasoning.” As Allie et al. define it, “point reasoning” reflects the view that the result of any trial could be, in principle, the correct measurement. Responses in categories R1, R2 and R3 clearly reflect this view. Not all the responses in R0 and R4 reflect this errant thinking, but R0 and R4 responses demonstrate a clear absence of “set reasoning,” and are therefore grouped with R1 through R3. Categorizing the responses in this way
substantially improves the inter-rater reliability, raising the agreement rate to over 90% for each pair of coders on both questions.

Instruction clearly has an impact on students’ responses to questions 1 and 2, but the results are disappointing, as Figure 4-2 shows. At UNC and in the SCALE-UP class at NCSU, where uncertainty analysis is taught, student understanding improves. The number of consistent set reasoners (i.e. students who gave “set reasoning” responses to both questions) increases while the number of consistent point reasoners decreases. Statistics suggest that it is unlikely that the difference between the pretest and posttest distributions for the UNC and SCALE-UP are due to random sampling. For these two groups, chi-square rejects the null hypothesis that the pretest and posttest distributions are identical ($X^2=19.3$, $p<0.0001$, df=2 for UNC and $X^2=6.0$, $p=0.05$, df=2 for SCALE-UP). The students enrolled in the NCSU traditional laboratory appear to lose ground, but the result is not statistically significant ($X^2=2.6$, $p=0.27$, df=2). The phrase “statistically significant” in this chapter implies an $\alpha$-level of 0.05. That means that “statistically significant” results have a likelihood of false positive due to sampling of less than 5%. Even though the gains at UNC and in SCALE-UP are statistically significant, they are disappointing. After instruction, less than 30% of each treatment group consistently expresses expert-like views concerning the purpose of multiple measurements.
4.1.2 Question 3: Inclusion or exclusion of an outlier

Question 3 presents students with a list of repeat measurements that includes an outlier. Students are asked to state whether the suspect data point should be included and to explain the reasoning behind their choice.

When the published categorization scheme for responses to question 3, shown in Table 4-2, was applied to the data from the present study, several difficulties arose. Not all responses fit into the four categories provided. Many responses fit into more than one category. In particular, students who chose to exclude the outlier often provided more than one reason for exclusion. Further, the categorization provided by Allie et al. fails to address at least one significant theme in the student responses from the present study. More than 30% of the responses express the concern that inclusion of the outlier results in an inaccurate estimate of the measurement. Another minor theme not addressed in Allie et al.’s scheme is the use of statistical tests to identify outliers. About 5% of the responses suggest the student is aware that a statistical test for outlier exists. In some instances, specific tests are mentioned, but there is no single popularly used test. Most of the stated
tests appear to be ad hoc, and almost all fail to compare the variability of the data set without the outlier to the variability of the data set including the outlier.

Table 4-2. Response categories for anomaly question.
This is the categorization scheme published by Allie et al (1998) for student responses to question 3.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN1</td>
<td>The anomaly must be included when taking an average since all readings must be used.</td>
</tr>
<tr>
<td>AN2</td>
<td>The anomaly is noted, but it has to be included since it is part of the spread of results.</td>
</tr>
<tr>
<td>AN3</td>
<td>The anomaly must be excluded as it is most likely a mistake.</td>
</tr>
<tr>
<td>AN4</td>
<td>The anomaly must be excluded as it is outside the acceptable range.</td>
</tr>
<tr>
<td>AN0</td>
<td>Not codeable.</td>
</tr>
</tbody>
</table>

Student performance on this question is excellent, as shown in Figure 4-3. Even before instruction, more than 75% of the students indicate that the outlier should not be included in the average, with more than a third of these students citing concern for accuracy of the mean. After instruction, nearly 85% of the students choose to omit the outlier, with about half of these students citing concern for accuracy of the mean. Only SCALE-UP shows a statistically significant gain at the $\alpha=0.05$ level ($p<0.002$ for $z=3.3$ and $N=126$).

![Fraction of students who exclude the outlier](image)

Figure 4-3. Pretest/posttest comparison for question 3.
Each bar represents the fraction of students who indicate that outlier should be excluded when calculating the average. Error bars indicate 95% confidence intervals.
The result is not surprising, since the topic was discussed in the SCALE-UP classes on several occasions. In the first lab of the semester, the SCALE-UP students measured the average mass of pennies as a function of year. A graph of the class’ data was constructed by combining data from all groups. When students viewed the graph, outliers caused by student mis-measurement were clearly visible. It is likely that the explicit discussion of the topic in class improved student understanding.

4.1.3 Question 4: Spread as a measure of data quality

In question 4, students are asked to compare two sets of repeated measurements. The data sets have the same mean, but different spreads. Students are asked to choose between the following opinions and defend their choice:

A: The data set with the smaller spread is of better quality.

B: The data sets are of equal quality because the means are equal.

A vast majority of the explanations offered in the student responses to this question are restatements of the position chosen and provide little insight into student thinking. As a result, only the letter choice is analyzed for this question. As seen in Figure 4-4, the fraction of students who support the expert-like opinion expressed in choice “A” increases with instruction for each group of students.

![Figure 4-4. Pretest/posttest comparison for question 4.](image)

Each bar represents the fraction of students who chose “A” on question 4. Choice A is consistent with the expert-like opinion that spread serves as a measure of data quality. Error bars on the graph represent 95% confidence intervals.
The increase is statistically significant for SCALE-UP (p<0.002 for z =3.1 and N=126) and for UNC (p<0.0001 for z =4.4 and N=222). There is a significant difference in pretest scores across the three groups. The cause is unclear. No group had relevant instruction prior to the pretest to explain the observed differences.

4.1.4 Question 5: Assessing agreement between sets of repeat measurements

In question 5, students compare two sets of repeated measurements. The data sets have numerically different means and similar spreads, but the two data sets have considerable overlap. Students are asked to choose between the following opinions and defend their choice:

A: The results agree.

B: The results don’t agree.

Both the choice and the reasoning provide substantial insight into student thinking for this question.

Even before instruction, most students defend the “results agree” position, as shown in Figure 4-5. Instruction appears to have little effect on the fraction of students who defend the expert position. Only the UNC group shows a statistically significant increase (p<0.001 for z=3.5 and N=217) in the fraction of students selecting the right answer and the difference is relatively small (from 65% correct on the pretest to 80% correct on the posttest).

Figure 4-5. Pretest/posttest comparison for question 5: Do results agree?
Each bar represents the fraction of students who state that the two data sets agree. Error bars represent 95% confidence intervals.
Most of the student responses from the present study fit easily into the categorization scheme from Allie’s South African study (shown in Table 4-3).

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMSS1</td>
<td>It depends on how close the averages are.</td>
</tr>
<tr>
<td>DMSS2</td>
<td>It depends solely on the relative spreads of the data</td>
</tr>
<tr>
<td>DMSS3</td>
<td>It depends on the degree of correspondence between individual measurements in the two sets</td>
</tr>
<tr>
<td>DMSS4</td>
<td>It depends on both the averages and the uncertainties</td>
</tr>
<tr>
<td>DMSS0</td>
<td>Not codeable</td>
</tr>
</tbody>
</table>

Table 4-3. Response categories for question 5.
This is the scheme published by Allie et al (1998) for student responses to a question concerning the agreement of two data sets.

In the present study, more than 80% of the explanations either judge agreement of the two data sets based solely on the comparison of the averages (DMSS1) or argue that the difference in the means can be explained by the uncertainty in the measurements (DMSS4). However, the remaining 20% of the responses were not easily classifiable as DMSS2 and DMSS3. In particular, several student responses argue that data sets agree/disagree because “ranges” are similar/different. However, what the student means by “range” is often unclear. Some students clearly use the word “range” to denote some spread measure of the data. As in these examples, this spread measure is most often the difference between the highest and lowest values:

Group A's range is 10 and group B's range is 10 so their averages would be similar to each other. [student paper 77]

They agree in one way which is that their range was both 10 so that they both had pretty good technique. Their averages do however differ. [student paper 108]

Others use “range” to describe the set of values between the maximum and minimum data values:

Though the first group had results ranging from 118-128 and the second had a range from 121-131 I would say that they both have about the same results. I would say that the results agree. [student paper 2]

The values for both groups where in a similar range between 118 and 131. And the results were only a 2mm difference. [student paper 25]

In many instances, the meaning cannot be gleaned from context:

Even though they are not precisely the same number, they both lie within the range of both experiment's results. [student paper 30]
C; neither one b/c the range of answers are fairly close and so is the average, so they don't agree, so if I had to choose one, it would be B. But for the variables than haven't been accounted for I would say they are pretty close [student paper 64]

Such responses may represent either DMSS2 thinking (novice-like thinking) or more sophisticated thinking linked to the use of the uncertainty overlap condition.

Before instruction, more than 80% of students in every group judge the agreement of the two data sets based solely on the comparison of the means, while fewer than 5% use uncertainty arguments to conclude that the two sets of data agree. Appropriate instruction changes this situation changes dramatically, as shown in Figure 4-6. In the traditional labs at NCSU, there is no instruction on uncertainty. After instruction, the overwhelming majority still bases their judgment of agreement on how close the means are without any reference to the uncertainty inherent in the measurements. In both the SCALE-UP class and the lab at UNC, the number of students who use uncertainty arguments increases dramatically as the number of “means only” reasoners drops.

![What do students use to assess agreement between data sets?](image)

**Figure 4-6. Pretest/posttest comparison: How students assess agreement on question 5.**
For students in SCALE-UP and at UNC, the fraction of students who use the concept of spread to assess the agreement between two sets of repeated measurements increases, decreasing the number of students who assess agreement based on an arbitrary determination of the size of the difference between the means. Error bars represent 95% confidence intervals.

While the UNC students made stronger gains on this question than the SCALE-UP students, instructional treatment at UNC cannot be deemed an unqualified success. Even after instruction, only about half of the UNC students use the concept of uncertainty to assess agreement of the two data sets.
4.1.5 **Question 6: Reading a ruler**

Students are asked to read the length of a bug in a picture using the ruler in the picture and choose between five alternatives of increasing resolution. Choices A through C are too conservative, representing 100% confidence intervals for the measurement. None of these choices requires interpolating between the markings on the ruler. Choice D is most in line with expert opinion as expressed by the ISO *Guide*, which advocates a realistic (rather than “safe”) evaluation of uncertainty. Choice D corresponds to estimating one decimal point beyond the marks on the ruler. Choice E is unrealistically restrictive. Students were asked to explain their choice.

Taken together, choices A and B account for fewer than 3% of all responses. Most of the explanations provided by these students indicated confusion about the intended definition of the bug’s length. Choices C and D were the most popular, accounting for about 90% of the papers. The explanations given by students who selected choices C and D generally focus on the process of reading a ruler or on the resolution of the ruler. Choice E accounts for the remaining papers. Most of these papers indicate that choice E was selected because it was the most exact answer. As seen in Figure 4-7, only SCALE-UP shows a statistically significant increase in the fraction of students selecting the correct answer (p<0.0001 for z=4.9 and N=125).

![Question 6: Reading a ruler](image)

**Figure 4-7. Pretest/post comparison for question 6.**
This shows the fraction of students who chose the two most popular answers to question 6. Choice D is the correct answer. Choice C corresponds to reading a ruler without interpolating between the markings on the ruler. Choice C is the most popular wrong answer. Error bars represent 95% confidence intervals.
SCALE-UP is also the only group that shows a statistically significant decrease in the most popular wrong answer (p<0.05 for z=-2.2 and N=125). Specific homework assignments in the SCALE-UP class involving reading a ruler may have contributed to the gain in the fraction of correct answer.

4.1.6 Question 7: Stacking

Question 7 is similar in format to question 6. Instead of a bug, a picture of a stack of floppy disks with a ruler is presented. Students are asked find the average thickness of floppy disks in the picture and choose between 5 alternatives of increasing resolution. The correct answer (choice D) is achieved by measuring the combined thickness of the ten disks in the picture, interpolating between the marks on the pictured ruler, and dividing by 10. If stacking is not employed and the ruler is read correctly, an answer consistent with choice C is achieved. Choice C can also be achieved if stacking is employed but the ruler is read conservatively (i.e. without interpolation). Choice B is consistent with measuring the thickness of a single disk without interpolation. Choice A is overly conservative. Choice E is unrealistically restrictive.

Student reasoning was solicited by two questions. The following question was posed in between the picture and the multiple choice question: “What method should be used to measure the average thickness of the floppy disks in the picture as accurately and precisely as possible?” Following the multiple choice question, students were asked to explain their answer.

The explanations provided substantial insight into whether the student employed stacking. About 40% of the papers indicate convincingly that the student measured the combined thickness of five or more disks and divided by the number of disks in the stack. Students who employed stacking left various types of evidence in the two explanations including a verbal description of the technique, a calculation for the average, marks on the picture indicating the two ends of the stack, or a measurement for the overall thickness of 10 all disks. About 10% of the papers mentioned the technique in passing, but provided no additional evidence that the technique was used, often suggesting that the technique be used. The remaining papers either describe the measurement of a single floppy or provide no evidence that stacking was employed. Figure 4-8 shows that instruction has a significant impact on whether students practice stacking.
Question 7: Do students employ stacking?

![Bar chart showing pretest/posttest comparison of stacking in question 7.](image)

**Figure 4-8. Pretest/posttest comparison of stacking in question 7.** This shows the fraction of students who employ stacking of more than 5 disks to determine the average thickness. Error bars represent 95% confidence intervals.

Both SCALE-UP and UNC show statistically significant gains in the fraction of students who employ stacking. For SCALE-UP, \( p<0.0001 \) for \( z=4.5 \) and \( N=122 \). For UNC, for \( p<0.0001 \) for \( z=3.7 \) and \( N=222 \).

Both groups did at least one lab activity in which stacking was employed. Perhaps the most striking feature is that the pretest score in the SCALE-UP class is significantly higher than the pretest scores of the other two groups. This is most likely also a treatment effect. In the hour before the pretest was administered in the SCALE-UP classes, one of the class activities was to estimate the thickness of a single page of their textbook. This single 10-minute activity (and its timing) clearly increased the fraction of students that employ stacking.

The distribution of multiple choice answers to question 7 is also informative. No student chose A on either the pretest or the posttest. Students universally feel that the average thickness of disks can be determined at least to within the smallest marking on the ruler. Choices B, C, and D account for over 90% of student choices, with each choice representing more than 20% of the papers. Figure 4-9 shows the fraction of each group choosing the correct answer. Only SCALE-UP shows a significant increase (\( p<0.0001 \) for \( z=4.7 \) and \( N=122 \)).
Figure 4-9. Pretest/posttest comparison of correct answers to question 7. This shows the fraction of students who chose the correct answer (Choice D). Error bars represent 95% confidence intervals.

Not all students who stack select the correct answer. Many fail to use the full resolution of the ruler. Others who stack round off the result, often asserting that the average value is limited to the level of resolution afforded by the measurement device.

On the posttest, the SCALE-UP students selected the correct answer because they used the correct techniques (stacking and interpolation between the marks on the ruler), as Figure 4-10 shows. This is in sharp contrast with both UNC and NCSU traditional students.

Figure 4-10. How stacking influences student answers to question 7.
These mosaic plots show which multiple choice answer students selected as a function of whether they employed stacking to find the average thickness of a floppy disk. The height (width) of each rectangle is proportional to the fraction of students that fall into each category along the vertical (horizontal) axis.

SCALE-UP students who employed stacking were far more likely to select the correct answer than their classmates who apparently did not use the technique. Conversely, very few of the students who failed to employ stacking chose the correct answer. This effect is far more pronounced for the SCALE-UP students than for the other groups. Half of the SCALE-UP students employed stacking to select the correct answer. More than 85% of the SCALE-UP students who selected the correct answer were stackers. At UNC, only 40% of the students employed stacking to select the correct answer, and only about 60% of the students who selected the correct choice did so using the correct measurement technique. (Notice that students who did not use stacking but selected the correct answer overstate the certainty of their measurement by a factor of ten!) In the NCSU traditional laboratories, very few students employed stacking, and most of those who did read the ruler without interpolation or rounded their answer off to the smallest mark on the ruler.

4.1.7 Analyzing the responses to items 8-11

On items 8-11, students were asked to report a measurement and an uncertainty estimate for that measurement. Students were also asked to explain how the measurement was made and to provide the reasoning behind the numbers that were reported. In general, students reported decimal numbers (with and without units) for the measurements and uncertainties. Analysis of these numerical responses focuses principally on issues related to student understanding of uncertainty. Special attention is paid to the values that students report for uncertainty and the number of significant figures students report for the measurements and uncertainties. Less attention is paid to the accuracy of student measurements and student regard and understanding of units of measure. The written descriptions of the measurement techniques and student reasoning are analyzed to determine what tools and techniques the students employed to make their measurements and to determine how the uncertainty was estimated.

The pretest results indicate that very few, if any, of the students had encountered the concept of uncertainty (in the ISO sense of the word) before instruction. Most students wrote numbers in the uncertainty blank on the pretest, but many did not. Among the numerical responses, there is astounding variety of
answers. For example, on question #8, where all students are provided with the same fictitious data to interpret, there are more than seventy distinct student numerical responses for the uncertainty. Only one of these numerical responses occurs on more than 10% of the papers. Students who did not write a number in the space provided often left the space blank, or wrote “I don’t know” or some equivalent. Others wrote verbal responses that reflect a lack of understanding of the concept of uncertainty. These responses often describe sources of uncertainty or express skepticism about the value of the measurement. The stunning cornucopia of incorrect responses suggests that these students don’t understand what uncertainty is before the pretest. Traditional NCSU students also reflect similar on the pretest.

4.1.8 Question 8: Removing an outlier to calculate a mean

Question #8 presents students with a list of values for a measurement. Students are asked to estimate the mass of a marble from a list of multiple measurements that includes an outlier. Students are also asked to estimate the uncertainty in the mass.

**Question #8: Recognizing and treating an outlier (AN2)**

8. Several students measure the mass of the same marble on a scale that can be read to the nearest 0.01 g. The results are as follows: 50.91 g, 50.90 g, 50.87 g, 58.03 g, 50.82 g. What value should they report for the mass of the marble? for the uncertainty in the mass?

   Mass = ___________

   Uncertainty in the mass = ___________

   Explain how you determined your answers.

A correct solution requires that the student recognize and remove the outlier. The mass may be estimated by the mean of the values (excluding the anomaly) and the uncertainty can be estimated by the standard deviation of the values (excluding the anomaly).

As in question 3, a large majority of students excludes the outlier, even on the pretest. The results are shown in Figure 4-11. The pretest/posttest difference is significant at the $\alpha=0.05$ level for both SCALE-UP and UNC. (For SCALE-UP, $p<0.02$ for $z=2.4$ and $N=127$. For UNC, $p<0.003$ for $z=3.2$ and $N=219$).
The number of students who exclude the outlier for this question is slightly lower than in question 3. About 80% of students treated the outlier the same way in both questions. Only about 10% of the students who indicate that the outlier in question 3 should be excluded fail to exclude the outlier in question 8. These students appear to have calculated the mean without critically examining the data and failed to notice that the mean they reported (about 52.3 g) does not resemble any of the numbers in the data set.

Despite the constrained nature of this problem, more than forty different values for the mass of the marble were reported by students. Most of these responses occur on fewer than 5 papers each. This diversity principally reflects the combined impact of arithmetic mistakes, choices concerning how to round off the number and inclusion/exclusion of the outlier. The seven most popular responses account for more than 80% of all papers, as shown in Table 4-4.

Table 4-4. Popular student responses for the mass of the marble in question 8.
The frequency represents the fraction of all papers with each response.
<table>
<thead>
<tr>
<th>Response</th>
<th>Frequency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.88 g</td>
<td>47%</td>
<td>Most correct answer. Mean, excluding outlier, rounded to 0.01 place.</td>
</tr>
<tr>
<td>52.31 g</td>
<td>14%</td>
<td>Mean, including outlier, rounded to 0.01 place.</td>
</tr>
<tr>
<td>50.87 g</td>
<td>9%</td>
<td>Acceptable answer. Either indicates the median (no rounding needed) or the mean, excluding outlier, incorrectly rounded to 0.01 place.</td>
</tr>
<tr>
<td>50.875 g</td>
<td>8%</td>
<td>Mean, excluding outlier, unrounded</td>
</tr>
<tr>
<td>52.306 g</td>
<td>4%</td>
<td>Mean, including outlier, unrounded</td>
</tr>
<tr>
<td>50.9 g</td>
<td>3%</td>
<td>Mean, excluding outlier, rounded to 0.1 place</td>
</tr>
<tr>
<td>52.3 g</td>
<td>3%</td>
<td>Mean, including outlier, rounded to 0.1 place</td>
</tr>
</tbody>
</table>

Three responses are considered correct: 50.88 g, 50.89 g, and 50.87 g. These are the mean, trimmed mean (i.e. the mean with the two most extreme values removed) and the median. The correct response must have four significant figures, since reasonable uncertainty estimates lie between 0.02 g and 0.05 g. The fraction of students giving the correct answer is shown in Figure 4-12.

![Question 8: Correct answer for mass](image)

**Figure 4-12. Pretest/posttest comparison for question 8: Correct mass?**
This shows the fraction of students who give a correct value for the mass. Three different answers (50.88 g, 50.87 g, and 50.89 g) are considered correct. Error bars represent 95% confidence intervals.

The pretest/posttest difference is statistically significant at $\alpha=0.05$ for both SCALE-UP ($p<0.05 \ z=2.2$ and $N=127$) and UNC ($p<0.0001$ for $z=5.5$ and $N=219$). The difference between the posttest scores of SCALE-UP and UNC is somewhat surprising, since the fraction of students who excluded the outlier from the calculation was about the same for both groups (see Figure 4-11). The difference, however, is largely explained by a
group of students in the SCALE-UP class who computed the mean without the outlier, but failed to round the
mass value to four significant figures. Very few of the UNC students did this on the posttest.

There are several ways to calculate a reasonable estimate for the uncertainty in the mass of the
marble. The most expert-like method employed by the students is to calculate the standard deviation of the
mean of the masses not including the mean. Other reasonable methods for making uncertainty estimates used
by the students include calculation of the standard deviation and average deviation. On the pretest, fewer than
10% of the students applied any of these expert-like methods. The most popular response on the pretest is 0.01
g, which occurs on about 20% of the pretest papers. Only three other responses (0.005 g, 0.05 g, 58.03 g)
 occur on more than 5% of the pretest papers. Clearly, many students based their uncertainty estimate on the
resolution of the scale.

Examination of student explanations on the pretest reveals that students used a wide variety of non-
expert-like methods to estimate the uncertainty of the mass of the marble. Some estimated uncertainty using
simple spread measures such as the difference between highest and lowest data value. Some of the spread
measures account for number of trials while others don’t. Other students performed calculations that assessed
the difference between two different ways of estimating the mass. For instance, some students calculated the
difference between the mean with the outlier included and the mean with the outlier excluded and reported the
result as the uncertainty. Still others multiply (or divide) the resolution of the scale by the number of trials. No
single method dominates and it is unclear that students have any confidence of the value of any of the
measures. These novice methods for estimating uncertainty may have been invented to fill the blank on this
test or may be the result of prior instruction.

The posttest tells a different story. Students at UNC and in the SCALE-UP class were taught how to
calculate standard deviation and standard deviation of the mean. Not surprisingly, uncertainty estimates based
on standard deviation emerge on the posttest for UNC and SCALE-UP. In sharp contrast, none of the NCSU
traditional group used either statistic on the pretest or the posttest. On the posttest, the standard deviation of
the mean (0.02 g) is the by far the most popular answer among UNC and SCALE-UP students, appearing on
about 42% of the UNC posttests and 36% of SCALE-UP posttests. The standard deviation of the masses (0.04
g) is also popular at UNC, appearing on about 12% of the posttests. However, the answer 0.01 g remains
popular, appearing on 13% of UNC posttests (16% for SCALE-UP). Examination of the student explanations confirms that students who chose the popular answers (0.01 g, 0.02 g, 0.04 g) did so for the expected reasons (resolution of the scale, standard deviation of the mean, standard deviation of the masses). It is interesting to note a mistake that some of the students (mainly at UNC) made. After miscalculating the standard error and getting a value that was too low (often as a result of failing to take the square root), some students argued that the resolution of the scale should be taken as the uncertainty because the standard error was smaller than the uncertainty due to the resolution of the scale. Sometimes students can demonstrate surprising understanding through their mistakes!

If only the answer 50.88 g with a standard uncertainty of 0.02 g is considered correct, about 40% of the UNC students (26% of SCALE-UP students) got the question correct on the posttest. If the definition of “correct” is extended to accommodate answers that use combinations of mean, trimmed mean, or median with alternative spread measures including standard deviation, and average deviation, the numbers improve to 49% correct at UNC (37% for SCALE-UP). While this is not universal success, this certainly is a remarkable improvement over the pretest (where less than 5% of each group met the either criteria of correctness). It is also much better than instruction that does not address the topic – none of the NCSU traditional lab students got the question right on the posttest.

4.1.9 Question 9: Measuring the diameter of a quarter

In question 9, students are asked to measure the diameter of a quarter coin as precisely and accurately as possible and to estimate the uncertainty of the measurement. All students had access to coins and rulers marked to 0.1 cm. Vernier calipers were also provided for the posttest at UNC, since students were taught how to use them during the semester. None of the SCALE-UP students or NCSU traditional lab students used calipers on either administration of the test. None of the UNC students used calipers on the pretest, but over 80% of the students at UNC used vernier calipers on the posttest. As a result, it is difficult to meaningfully use the posttest results of UNC students in comparisons concerning students’ use of rulers. Immediately prior to the posttest, several UNC teaching assistants provided specific instruction on reading vernier calipers, further invalidating the UNC posttest results.
A quick desktop measurement using a ruler with mm markings gives 2.41 cm with an estimated standard uncertainty of 0.03 cm. The uncertainty is a “seat of the pants” estimate of a 70% confidence interval for the measurement. The estimate is one third of the distance between markings on the ruler. The vernier calipers used at UNC improve the quality of the measurement somewhat, reducing the (instrumental) uncertainty to about 0.005 cm. The uncertainty of the ruler measurement might also be decreased somewhat by “stacking,” but it’s unclear how much benefit the method gives, considering the practical difficulties keeping the quarters lined up and in contact while taking the measurement. Since only two students employed stacking, the method will not be discussed.

Despite the apparent simplicity of this question, many students made mistakes that complicate the analysis of this question. Some students fail to report the units of the measurement (13% of all responses). Others appear to report the wrong units for their reading (7% of all responses). In many cases, the student’s intent is easily understood; the vast majority (about 90%) of student responses were numbers between 2 and 3 (which were interpreted as diameter in cm, regardless of the unit supplied by the student) or numbers between 20 and 30 (which were interpreted as diameter in mm). However, numbers outside these two ranges occur among the responses. Examination failed to uncover themes among these papers, since the verbal descriptions that accompanied responses outside the two ranges rarely provided adequate information for determining the student’s intent. Many students failed to provide units for the uncertainty. Unless the student indicated otherwise, it was assumed that the uncertainty and the diameter value were given the same units.

For students who used a ruler, two forms of answer dominate the responses for the diameter of the quarter. Some students make mistake-free readings of the ruler and interpolate between marks on the ruler. Students who do this report diameters of the form 2.XX (cm) or 2X.X (mm). Students who make mistake-free, but conservative, measurements report diameters of the form 2.X (cm) or 2X (mm). The results, shown in Figure 4-12, mirror the results of question 6 (the “ruler-bug” question).
Measuring the diameter of a quarter: Do students interpolate?

Pre Post Pre Post

Fraction of students

2.X cm 2.XX cm

NCSU Trad. NCSU SCALE-UP

Figure 4-13. Pretest/posttest comparison for question 9.
This shows the fraction of students reporting one of the two most popular responses. Responses of the form 2.XX cm (or equivalent) indicate that the student interpolated between the marks on the ruler. Responses of the form 2.X cm (or equivalent) indicate that the student only read to the marks on the ruler. Error bars represent 95% confidence intervals. UNC results are excluded because 80% of the UNC students used vernier calipers on the posttest.

The NCSU traditional lab group shows no change in response patterns, while SCALE-UP shows a shift from novice to expert practice. The results for UNC are not included on Figure 4-13 because very few UNC students used a ruler for this measurement on the posttest. It is interesting to note, for reference, that the pretest performance of UNC students on this question is better than either group at NCSU, with 43% of students giving 2.XX type answers and 32% giving 2.X type answers. It is also interesting to note that about 20% of all student responses do not fall into either of these broad classes. These remaining responses are mainly populated by values for diameters quoted to more than 3 significant figures and values for diameter that are wildly inaccurate.

Histograms of the values of diameter revealed two striking observations about student measurement practices. A histogram of the 2.X type answers (Figure 4-14) shows that the students who choose to report only to the marks on the ruler often take sloppy measurements. Only half of these students get the right value (2.4 cm). While it is somewhat encouraging that most of the remaining students in the 2.X group are within 1 mm of the correct value, a few seconds with a ruler and quarter should convince an expert ruler user that mis-measurement by 1 mm represents extremely sloppy placement of ruler and coin.
Figure 4-14. Student measurements: histogram #1.
Distribution of diameter values reported to two significant figures. Both pretest and posttest results are shown.

A histogram of the 2.XX type answers (Figure 2-15) reveals two sharp spikes at 2.35 and 2.45 superimposed on a bell shaped mound centered near the 2.41 mark.

Figure 4-15. Student measurements: histogram #2.
Distribution of diameter values reported to three significant figures. Pretest and posttest results are combined.

The distribution clearly includes some students reading to the nearest half-marking on the ruler and others who are interpolating more closely. It’s unclear what the absence of a similar spike at 2.40 means. Perhaps students who read to the nearest half mark only report the extra significant figure when the value falls between
two marks on the ruler. In any event, the presence of the spikes “on the fives” shows that number of students reporting 2.XX values is likely to exceed the number of students doing careful interpolation.

While the fact that large numbers of UNC students used calipers on the posttest renders many comparisons invalid, some comparisons are possible. The most striking feature of the UNC posttest is the large numbers of UNC students measured the diameter multiple times. Almost a third (32%) use either standard deviation or standard deviation of the mean to estimate the uncertainty in the measurement. Given the nature of the measurement, this seems like a waste of time. Since the UNC students were taking the posttest for a grade, many may have taken multiple measurements simply to demonstrate that they use statistics to calculate uncertainty. By contrast, only about 10% of the SCALE-UP students used the spread of multiple measurements to estimate uncertainty. It may be that the SCALE-UP students were lazier (the posttest was graded in SCALE-UP, but counted only for extra credit), or it may be that the SCALE-UP students recognized that multiple trials are overkill for this measurement.

It is also clear that a large fraction of the UNC students who used calipers for the posttest did not take full advantage of the measurement tool. The calipers the students used have least count markings of 0.005 cm, yet about 33% of the students who used the calipers reported answer of the 2.XX (or 2X.X) form. About 47% report answers of the 2.XXX (or 2X.XX) form. However, a closer look at these answers, shown in Figure 4-16 reveals a disproportionate number ending in either zero or five.

![Figure 4-16. Histogram of the last digit in diameter value.](image)

Distribution of last digit of diameters reported to four significant figures by students who used calipers to measure the diameter of a quarter.
As with the ruler, an additional significant digit in the result does not always mean that the student interpolated when making the reading.

4.1.10 Question 10: Measuring the mass of a paper clip

For question 10, students were asked to determine the mass of a typical #3 Gem brand paper clip as accurately and precisely as possible. Blanks for the mass, for uncertainty, for describing the measurement process and for explaining how the uncertainty was determined were provided. Although the question does not ask the student to do so, stacking is clearly indicated for this measurement. The variability among paper clips is likely small compared to the uncertainty in the mass due to the scale. A simple estimate of the uncertainty in the average mass of a sample can be obtained by dividing the “seat of the pants” estimate of the 70% confidence interval in the total mass by the number of paper clips in the sample. At NCSU, students used a triple beam balance with marks 0.1 g apart. Students could easily interpolate between markings on these scales to get measurements with standard uncertainties on the order of 0.02 g. The scale is tared by turning a screw. At UNC, students used electric pan balances. The digital displays have a least count of 0.1 g and the scales are tared by the push of a button.

In almost all cases, the descriptions provided adequate information to determine whether the student employed stacking or not. From many descriptions, there was little doubt that stacking was employed. In these explicit descriptions, students wrote total mass, the number of clips placed on the scale at once, division calculations provide conclusive evidence. Other descriptions gave convincing evidence that the student failed to stack by identifying some other method. Most common among these are multiple trials of one clip, single measurements of a single clip, and methods that involved a mixture of stacking and multiple trials.

The results of this question mirror the results of the floppy disks question (Question #7). Instruction on stacking clearly had a large impact on whether students employed stacking. Figure 4-17 shows the fraction of each group with students employing stacking with 5 or more paper clips.
Notice that the fraction of the SCALE-UP students who stack on the pretest is at least as large as the posttest fraction of stackers in the UNC and NCSU traditional groups. A single, ten minute long activity (measuring the thickness of a page in the textbook) done by the SCALE-UP students in the hour before the pretest was administered appears to have been at least as effective as an entire semester of lab work in the other two treatments. Subsequent activities in the SCALE-UP classroom added substantially to the gain. The gain posted by the UNC group is statistically significant (p<0.0001 for z=5.5 and N=220), but small. Some of the students who employed stacking also took multiple measurements, often in an attempt to evaluate the uncertainty of the measurement using statistics.

Students who stacked had difficulty making reasonable uncertainty estimates on this question. Only a fraction of the students who stacked estimated the uncertainty by dividing the resolution of the scale by the number of clips in the sample. The SCALE-UP group performed best in this regard, with about a third of the SCALE-UP posttest stackers (36 of 109) using this method to estimate the uncertainty. At UNC, only 11 of the 59 posttest stackers use this method. By contrast, 27 of the 59 posttest stackers at UNC estimated the uncertainty of the measurement based solely on the resolution of the scale. Estimating the uncertainty in this way ignores the fundamental reason for stacking. Among SCALE-UP posttest stackers, this was somewhat less common (14 of 109). It should be noted that, for a substantial fraction of students (~30%), it is difficult to
classify the method used to estimate uncertainty for this measurement. In many cases, student explanations do not provide enough information about how the value was found. In other cases, the method used is unique. With few exceptions, these unclassifiable responses fail to reflect expert-like thinking.

Non-stackers fell into two major groups: students who measured a single clip once and students who performed multiple trials, either with a single clip or with a different clip each time. Figure 4-18 shows the fraction of each treatment group in these two categories.

![Figure 4-18. Massing a single paper clip.](image)

Some students took several trials; others measured a single paper clip once. Error bars represent 95% confidence intervals.

By the posttest, very few UNC and SCALE-UP students remain in the “one clip once” camp. The students who chose multiple trials on the posttest at UNC are particularly interesting. Because the variability in the mass of paper clips is far less than the 0.1 g least count of the digital balance the UNC students used, students usually saw a string of identical measurements, with perhaps one value straying by 0.1 g. One would hope that, in the face of this evidence, that the students might quickly recognize that the measurement is clearly limited by the resolution of the scale, yet there is a group of students that mindlessly calculates mean and standard deviation from this list of numbers. While many of this group rounds the result to the least count if the scale, about 10% of the UNC students calculate a mean mass from multiple measurements of single clip samples and report the result to two significant figures. Most of this same group calculates a standard deviation or standard error to estimate the uncertainty.
4.1.11 Question 11: Timing the fall of a coffee filter

In question 11, students are asked to measure the time it takes a coffee filter to fall 1 m as accurately and precisely as possible. Blanks for the time, for uncertainty, for describing the measurement process and for explaining how the uncertainty was determined were provided. In all administrations of the test, stations were set up with metersticks taped flush with the floor, digital stopclocks, and coffee filters. All the stopwatches were identical digital stopclocks with least count of 0.01 s. Multiple trials are clearly indicated for this measurement, since the variability in the stopclock readings is likely to exceed the least count of the stopclock.

There is tremendous variation in the completeness of the descriptions that students provide for this question. In some cases, student notes are complete enough to determine what data was taken (and, in some cases, what data was excluded). In some cases, line-by-line calculations for both the time and its uncertainty are shown. Other students assert that certain measurements and calculations were made, but there is no data recorded or calculation shown on the papers. Still other students give explanations that offer little or no insight into how the results were obtained. The quality of the explanations is clearly affected by the stake the student has in the paper. The explanations provided by UNC students on the posttest are by far the most detailed, with 95% of all papers containing data. On the pretest, about 75% of the UNC students recorded their data on the page. Data appears on about 75% of the SCALE-UP papers and about 50% of the NCSU traditional papers. At NCSU, there is little difference between the pretests and posttests on the fraction of students reporting data.

Among the UNC and SCALE-UP students who report data, there is an interesting change in student practice. Before instruction, three appears to be the minimum acceptable number of trials to take. Less than 5% of the students who report data report fewer than three trials. About a third of the students who report data on the pretest report three data points, with the remainder reporting more than three trials. On the posttest at UNC and in SCALE-UP, the minimum acceptable number of trials rises to five. At UNC and in SCALE-UP, fewer than 5% of students reporting data report fewer than five trials. On the posttest, five trials is extremely popular, accounting for about two thirds of students who report data.

The methods that students use to estimate uncertainty on this test item are similar to the methods students report using for question 8 (the mass of the marble). On the posttest, most students use the ad hoc
methods to estimate uncertainty described in the analysis of question 8. As shown in Figure 4-19, the fraction of students at UNC and in SCALE-UP who use standard deviation and standard deviation of the mean increases dramatically as the fraction of students using novice methods decreases. The NCSU traditional students are excluded from Figure 4-19. No student in this group uses either statistic to estimate uncertainty on either the pretest or the posttest.

![Figure 4-19. Pretest/posttest comparison: Use of standard deviation.](image)

Fraction of students from each group that use standard deviation or standard deviation of the mean. None of the NCSU traditional lab students used these statistics on either the posttest or the pretest. Error bars represent 95% confidence intervals.

Before instruction, the two most popular identifiable methods for estimating are using the least count of the stopwatch as the basis for the estimate and using the difference between the highest and lowest data points as the basis for the estimate. The fractions of students employing these novice methods to estimate uncertainty are shown in Figure 4-20.
Both of these novice methods decline in popularity at UNC and in SCALE-UP, while the fraction of students estimating the uncertainty of the measurement based solely on the least count of the stopwatch increases for the NCSU traditional lab group.

The fraction of students who get reasonable answers supported by correct reasoning is disappointing for this question. For this question, an answer is deemed reasonable if it is in one of two forms: $1.XX \pm 0.0X$ or $1.X \pm 0.X$. The reasoning is deemed correct if the uncertainty was found by standard deviation, average deviation or standard deviation of the mean. Using these definitions, only about a third of the SCALE-UP students and about two thirds of the UNC students produce reasonable answers. While these represent strong gains over the pretest (4% of the UNC students and none of the SCALE-UP students get it “right” on the pretest), there is still a long way to go. It is interesting to note that of the 49 SCALE-UP students who used appropriate statistics on the posttest, only four fail to produce reasonable answers. At UNC, about 25% of the students of the students who compute the appropriate statistics fail to produce reasonable answers. SCALE-
UP students were taught to use a spreadsheet to compute standard deviation and standard deviation of the mean and had ample access to computers during the posttest. This was not the case at UNC.

### 4.1.12 Section summary: Pretest/posttest differences

Instruction appears to have a considerable effect on student learning. In the NCSU traditional labs, where uncertainty analysis and measurement are largely ignored, no statistically significant differences between pretest and posttest performance are found. In SCALE-UP and UNC, student performance on many items on the test is significantly better on the posttest. The SCALE-UP and UNC students appear to learn much about measurement and uncertainty, many of these students retain novice ideas and practices after instruction. This attests to the validity of the test. Further evidence of the validity of the test is that differences in performance on specific questions (e.g. questions 7 and 10) can be linked to differences in instruction.

### 4.2 Are different modes of thinking related?

The remainder of this chapter is devoted to uncovering possible relationships between modes of thinking. Some central questions are:

- Does the point/set classification scheme correlate with ruler reading practices?
- Does the point/set classification scheme correlate with proclivity for stacking?
- Are students’ ruler reading practices correlated with student proclivity for stacking?

The following sections explore these connections with the help of mosaic plots and inferential statistics for correlations between categorical variables. Unless otherwise stated, the results presented in the following sections are for the pretest only. Instruction, particularly in SCALE-UP and at UNC, is likely to have a noticeable effect on the nature of connections between concepts. Before instruction, the students are likely to exhibit “natural” thinking patterns (i.e. patterns which do not reflect the influence of a specific curriculum). Students at UNC and in SCALE-UP had grade motivation for giving correct answers on the posttest. As a result, students’ responses will likely reflect what the students feel the teacher wants to hear, rather than what the students are actually thinking.

Throughout this section, the chi-square test will be used to test the null hypothesis that the observed relationship between two categorical variables is due to chance. It is important to bear in mind that a
“statistically significant” chi-square result does not mean that the underlying relationship is strong. In fact, the large sample used in this study may be able to detect weak relationships. For this reason, measures of the strength of the association between the two variables will also be reported for the relationships explored in this section. Phi-square (chi-square divided by the number of data points) is one such measure of association. Phi-square varies from zero (indicating no relationship) to one (indicating one-to-one correspondence). The value can be interpreted as a “percent reduction in error.” A phi-square of 0.5 means that 50% fewer mistakes will be made if the prediction of the dependent variable is based on knowledge of the dependent variable than if the prediction is based solely on the knowledge of the distribution of dependent variable values.

Before exploring the relationships between different aspects of student thinking, the internal consistency of student performance on related test items will be assessed. These results provide a baseline for comparison when the strength of the relationship between different modes of thinking is explored. The results also provide a measure of the validity of the constructs used.

4.2.1 Internal consistency: Reading a ruler

Student ruler reading practices are assessed in questions 6 (ruler-bug), 7 (floppies), and 9 (diameter of the quarter). As noted earlier in this chapter, students choose to read a ruler in a variety of ways. Most students either read to the marks on the ruler or interpolate to one decimal place beyond the marks provided on the ruler. A small fraction of students report meaningless digits, grossly overstating the certainty of the measurement. An even smaller group of students underestimates the certainty of the measurement by reporting fewer significant digits than what is afforded by the markings on the ruler. In question 6, student ruler reading practice can be inferred directly from the student’s letter choice. Table 4-5 summarizes how the ruler reading practice is inferred from the student response to question 6.
Table 4-5. Inferring ruler reading practice from student’s letter choice for question 6.

<table>
<thead>
<tr>
<th>Letter Choice</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Inference not possible. Student may be confused about definition of bug’s length.</td>
</tr>
<tr>
<td>B</td>
<td>Student reads ruler (extremely) conservatively.</td>
</tr>
<tr>
<td>C</td>
<td>Student reads ruler conservatively</td>
</tr>
<tr>
<td>D</td>
<td>Student interpolates when reading ruler</td>
</tr>
<tr>
<td>E</td>
<td>Student overstates certainty of measurement</td>
</tr>
</tbody>
</table>

Student ruler reading practice can also be inferred from the response to question 7, but the inference must take into account both the student’s letter choice and whether stacking is employed. Table 4-6 summarizes how the ruler reading practice is inferred from the student response to question 7.

Table 4-6. Inferring ruler reading practice from question 7.
Student ruler reading practice on this question must be inferred from the letter choice and evidence concerning the student’s use of stacking. The one student who chose “A” for question 7 is not included in this analysis.

<table>
<thead>
<tr>
<th>Letter Choice</th>
<th>Evidence of stacking?</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>irrelevant</td>
<td>Student reads ruler conservatively (i.e. without interpolation)</td>
</tr>
<tr>
<td>C</td>
<td>yes</td>
<td>Student reads ruler conservatively</td>
</tr>
<tr>
<td>C</td>
<td>no</td>
<td>Student interpolates when reading ruler</td>
</tr>
<tr>
<td>D</td>
<td>yes</td>
<td>Student interpolates when reading ruler</td>
</tr>
<tr>
<td>D</td>
<td>no</td>
<td>Inference cannot be drawn about ruler reading</td>
</tr>
<tr>
<td>C or D</td>
<td>stacking mentioned, but no work shown</td>
<td>Inference cannot be drawn about student ruler reading practice</td>
</tr>
<tr>
<td>E</td>
<td>irrelevant</td>
<td>Student overstates certainty of measurement</td>
</tr>
</tbody>
</table>

The inferences about ruler reading drawn from these two questions are strongly correlated, as shown Figure 4-21. Qualitatively, the results make sense: students who fall in one category for question 6 are most likely to fall into the same category for question 7—that is, students who interpolate on question 6 are far more likely to interpolate on question 7 than students who read a ruler conservatively or students who overstate the certainty of the measurement. Similarly, a vast majority of students who interpolate on question 6 also interpolate on question 7. The number of students who read conservatively on one question and interpolate on the other is relatively small.
Figure 4-21. Ruler reading practice on questions 6 and 7.
This mosaic plot compares student ruler reading practices on question 6 (ruler-bug) to ruler reading on question 7 (floppies). Students who only read to the marks on the ruler (“conservative” ruler readers) on question 6 are much more likely to read the ruler conservatively on question 7 than students who interpolate between the marks or students who overstate the certainty of the measurement in question 7. On question 7, it was not always possible to infer student ruler reading practice from the student response. For this plot, N=585 pretest papers.

A chi-square test for independence clearly rejects the possibility that this relationship is due to sampling: $p<0.0001$ for $X^2=306$ and df=6. This relationship is quite strong: $\phi^2=0.52$. Knowing a student’s response to question 6 improves your chance of predicting the student’s ruler reading practice on question 7 by about a factor of two.

Ruler reading practice can be inferred from the response to question 9 (measuring the diameter of a quarter) by examining the number of significant figures reported for the diameter. On the pretest, all students were provided rulers with millimeter markings. Students who report values with one or two significant figures are deemed “conservative” ruler readers. Students who report three significant figures are deemed “interpolaters.” Students report four or more overstate the certainty of the measurement. The mosaic plot shown in Figure 4-22 shows that student ruler reading practice on this item is clearly related to student practices on question 6. Students who read the ruler conservatively on the ruler-bug question tend to report the diameter of the quarter in a manner consistent with conservative ruler reading, but the relationship is far less pronounced. The area on the graph that represents consistent practice decreases. The area that represents students who consistently overstate the certainty of the measurement drops drastically.
The statistics support the qualitative analysis. The association is statistically significant \((p<0.0001\) for \(X^2=55\) and \(df=6)\), but weak \((\phi^2=0.09)\). Student performance on the quarter measurement with ruler reading also correlates with ruler reading practice on the floppies question, as seen in Fig 4-23.
Again, the relationship is weak ($\phi^2=0.06$) but statistically significant ($p<0.0001$ for $X^2=36.5$ and $df=6$). Notice that the correlations cited in this paragraph are weak compared to the correlation between ruler reading practice on items 6 and 7. Making a real world measurement is a different task than making a measurement from a picture and then selecting the correct choice answer from a list. The multiple choice format cuts down on the diversity of student responses and provides students with clues about what the questions are intending to test. Further, there is little guarantee that knowledge translates to proper practice during the performance and reporting of an actual measurement. It is encouraging to note that correlation between questions 6 and 9 is considerably higher on the posttest than the pretest ($\phi^2=0.17$ vs. $\phi^2=0.09$). This suggests that the more students know about measurement, the more closely real world measurement practice reflects the reasoning the student uses on multiple choice items like the ruler-bug question.

4.2.2 Internal consistency: Stacking

Students who employed stacking on questions 7 (floppies) and 10 (measuring the mass of a paper clip) left considerable evidence to indicate they use the method. Students are considered stackers if more than five disks (or clips). As Figure 4-24 shows, students show consistent stacking behavior – students who employ stacking on the floppies question are more likely to stack when measuring the mass of the clip.
Figure 4-24. Stacking on questions 7 and 10.

This mosaic plot compares student use of stacking on question 7 (floppies) to student use of stacking on question 10 (mass of a paper clip). SCALE-UP students are not included in this comparison, since SCALE-UP students were exposed to stacking in the hour before the pretest was given. N=305 pretest papers.

SCALE-UP students are not included in Figure 4-24, since they received some instruction on stacking before the pretest. The results are statistically significant ($p<0.0003$ for $X^2=21.2$ and df=4), but weak ($\phi^2=0.05$). The weakness of this relationship is likely due to a ceiling effect. Many of the UNC and NCSU traditional students failed to stack on the pretest. A model that simply predicts that a certain small percentage of students will stack for the paper clip measurement will correctly predict the behavior of most students. Knowledge of their use of stacking on the floppies question will certainly improve the number of correct predictions, but only slightly. On the posttest, when students more students answer both questions correctly, the results show a much stronger relationship ($\phi^2=0.15$).

4.2.3 Internal consistency: Outlier exclusion

In questions 3 and 8, students must decide which data to include when computing an average. In question 3, students are asked directly whether a specific number (an obvious outlier) should be included in the average. In question 8, students are presented with a list of numbers and asked to report the mean. For this item, the student must recognize the outlier and discard it. As shown in Figure 4-25, student performance on these two items is highly consistent.
Figure 4-25. Outlier exclusion on questions 3 and 8.
In question 3, students are asked whether to include an obvious outlier when calculating an average. In question 8, students must identify an outlier and remove it to calculate an average. N=578 pretests.

The relationship is statistically significant (p<0.0001 for \(X^2=145\) and df=1) and strong (\(\varphi^2=0.25\)). Very few students who recommend excluding the outlier in question 3 fail to recognize and exclude the outlier in question 8. Even fewer students say that the outlier in 3 should be included in the average and then exclude the outlier in the calculation for question 8.

### 4.2.4 Internal consistency: Meaning of spread

Questions 4 and 5 probe student understanding of spread. In question 4, students compare two data sets of repeated measurements with the same mean but different spreads. They are asked to choose between two alternatives: A) the data set with the smaller spread is of higher quality and B) the data are of the same quality. Question 5 presents two sets of repeated measurements with slightly different means and equal spreads. The numbers are chosen so that there is considerable overlap between the two data sets. Students are asked to choose between two conclusions: A) the data sets agree and B) the data sets don’t agree. Students are also prompted to supply the reasoning behind their position. While there is no relationship between which letter the student chooses for the two questions, there is a relationship between the answer for question 4 and the reasoning given for question 5. The results are summarized in Figure 4-26.
Figure 4-26. Advanced set reasoning on questions 4 and 5.
This mosaic plot shows how student reasoning for question 5 is related to the answer chosen for question 4. Students who view spread as a measure of data quality are more likely to use spread when assessing the agreement of two sets of repeated measurements. N=590 pretests.

On question 4, some students argue that the data set with a smaller spread has higher quality, while a somewhat smaller group insists that spread is not a measure of data quality, arguing that the mean is of primary importance. Among this latter group of novice-like thinkers, almost all consider only the means when assessing the agreement of two data sets on question 5. Only a tiny fraction considers the spread of data in any way. Among students who support the more expert-like opinion on question 4, performance on question 5 is slightly better. A small group of these student argue that the data ranges overlap. Another small group argues that the difference between the means is small compared to some measure of spread. Yet, most of the students, regardless of their answer to question 4, fail to consider the spread when assessing agreement of two data sets on question 5. Despite this noticeable ceiling effect, the relationship is statistically significant (p<0.006 for $X^2=12.4$ and df=3), albeit weak ($\phi^2=0.02$). Posttest results show a stronger relationship ($\phi^2=0.05$). On the posttest, the frequency of correct answers and reasoning increases for both questions, but many students are inconsistent. Only 40% of students give expert-like answers on both questions, while about 15% give novice-like answers to both. On the posttest, most of the inconsistent thinkers (about 35% of all students) fail to use the concept of uncertainty when assessing agreement after answering question 4 correctly. More troubling is the group (about 5% of all students) who use the uncertainty condition to assess agreement.
and yet assert that spread is not a measure of data quality. These students are probably applying the uncertainty overlap condition in a rote way.

4.2.5 Is point/set reasoning related to other modes of thinking?

Student responses to questions 1 and 2 can be combined to characterize each student as “set reasoner,” “point reasoner,” or “transitional.” (For a description of these categories and procedures, please read section 4.1.1). This categorization does not appear to correlate with other modes of student thinking on either the pretest or the posttest. Chi-square tests for a relationship between point/set reasoning and each of the following variables yield no statistically significant results at the $\alpha=0.05$ level:

- ruler reading practice on questions 6, 7, and 9
- employment of stacking on questions 7 and 10,
- outlier exclusion on questions 3 and 8, and
- responses to questions 4 and 5 concerning the meaning of spread in data.

The bottom line is that the point/set classification scheme is not a good predictor of other aspects of test performance.

4.2.6 Is ruler reading practice related to stacking?

Chi-square tests between stacking performance and the measures of student ruler reading practice suggest that these two modes of performance are largely unrelated. Testing stacking practice on questions 7 and 10 against the three measures of ruler reading practice described in section 4.2.1 yields a few statistically significant results at the $\alpha=0.05$ level. Only five of the six possible pairs of variables give meaningful tests. (Testing ruler practice on question 7 against stacking practice on 7 has no meaning, since ruler reading practice was inferred based on student stacking practice). For the pretest, only two tests yield statistically significant, but weak results ($\phi^2<0.025$). For the posttest, the number of positive results increases to four of the five possible tests, but the relationships remain weak ($\phi^2<0.03$). Examination of the mosaic plots suggests an explanation. The small group of students that grossly overstates the certainty when using a ruler contains very few stackers. This group stands out clearly on two of the posttest plots and one of the posttest plots. There seems to be little difference between conservative and interpolating ruler readers concerning stacking.
4.2.7 Other related modes of thinking

For thoroughness, all ten possible pairs of the five variables (point/set reasoning, ruler reading, stacking, outlier exclusion and advanced set reasoning) can be tested. However, when many ad hoc comparisons are run, false positives are likely. For instance, twenty tests at an $\alpha$-level of 0.05 will yield, on average, one false positive. While chi-square tests reveal a few results that are statistically significant at $\alpha$-levels as low as 0.005, mosaic plots no consistent pattern among the constructs listed above. The only consistent relationship uncovered by the multitude of comparisons is between question 4 (concerning the meaning of spread) and outlier treatment on questions 3 and 8. The degree of association is weak ($\phi^2<0.04$) on both the pretest and the posttest, but in the expected direction—expert-like understanding of the meaning of spread is associated with slightly improved performance on questions involving outliers. No such pattern exists for question 5 (assessing agreement of two data sets) and questions concerning outliers.

4.2.8 Section summary: Are different modes of thinking related?

In this section, the degree of association between modes of student performance is assessed. On questions covering similar performance objectives (such items related to ruler reading), statistically significant relationships are found on the pretest. The results show internal consistency. For instance, conservative ruler reading practice on one item correlates to conservative ruler reading on another. In general, the strength of the relationship within related aspects of student performance increased from pretest to posttest, suggesting that instruction improved the connection between different measures of the same aspect of student performance.

When performance between different aspects of performance is compared (such as point/set reasoning and ruler reading), very few statistically significant results are found. When a statistically significant relationship between two questions is found, the results are universally weak compared to items measuring the same aspect of performance. Comparison between multiple questions within pairs of performance aspects (e.g. two measures of ruler reading and two measures of outlier exclusion) generally fails to give consistent behavior.
Chapter 5: Discussion

5.1 Overview

This project had two main objectives. The first was to investigate the impact of instruction on student understanding of measurement and uncertainty. The second was to identify patterns of student reasoning about measurement practices and measurement uncertainty. Most of the findings address the two research questions posed at the study’s outset:

- How does student understanding of measurement change over one semester from participating in an introductory physics course with a laboratory? How does this compare across different introductory physics laboratory curricula?
- What patterns among student practices and reasoning exist across various measurement tasks prior to instruction?

Findings of this study that pertain directly to these questions will be presented and discussed. Implications for teaching and research will be addressed.

5.2 Impact of instruction on student understanding

The most important finding of this study is that directed instruction on measurement practice and uncertainty calculation leads to improvement in student understanding. At UNC and in SCALE-UP, where measurement practice and uncertainty calculation were explicitly taught, student understanding improved significantly. Students in SCALE-UP and at UNC made strong gains concerning the purpose of multiple measurements, uncertainty estimation and the use of uncertainty to compare sets of data. In both laboratory programs, students were required to calculate uncertainty from multiple measurements and use the result to assess agreement of experimental results. In SCALE-UP, specific lab procedures were rarely provided. SCALE-UP students had to make choices concerning measurement. This may well have some impact on student measurement practices. The SCALE-UP group shows marked improvement in ruler reading practice and employment of stacking in appropriate contexts. At UNC, where lab procedures are largely provided for the students, there is no change in ruler reading practice and only modest gain in stacking. In the traditional
lab sections at NCSU, there is no instruction on measurement technique or uncertainty. The lab manual routinely asks the students to compare results without reference to uncertainty. Consequently, students learn nothing about uncertainty or its role in evaluating experimental results. The bottom line is that directed instruction yields results.

Despite marked improvement by the SCALE-UP and UNC students, the posttest results indicate that neither of these treatments is sufficient to propel all students to an expert level of understanding. Results are particularly disappointing concerning the purpose of multiple measurements. Even after instruction, only 30% of UNC students (20% of SCALE-UP) spontaneously identify construction of the mean or spread as a purpose of multiple measurements. Of these, most mention construction of the mean as the main purpose and only a tiny fraction identifies the quantification of spread as a purpose of multiple measurements. Students in both groups learn to estimate uncertainty, but it is unclear that they universally understand the purpose behind the exercise. Fewer than 60% of UNC students (30% of SCALE-UP) use the uncertainty when assessing agreement between two sets of repeated measurements. While these represent substantial gains over the pretest, the final state is somewhat disappointing.

There are also indications some of the student performance on the posttest may reflect rote application of procedures, rather than true understanding. The most striking example is the performance of UNC students on the paper clip question. Despite using a digital balance with a least count of 0.1 g to measure the mass of a 0.3 g paper clip, about 70% of the UNC students failed to employ stacking. Instead, they wasted time making meaningless repeated measurements of a single paper clip in order to be able to calculate a standard deviation. Another example concerns the meaning and use of uncertainty. After instruction, many UNC and SCALE-UP students use uncertainty when assessing agreement of two sets but argue that spread is not a measure of data quality. Such inconsistent behavior suggests that students have failed to develop a robust and coherent understanding of uncertainty.

This study also shows that a single ten minute activity can have a tremendous impact on student understanding. In the hour before the pretest was administered in the SCALE-UP class, students estimated the thickness of a page in the textbook with a ruler in an activity called “Sizing up the book.” The impact is astounding. SCALE-UP pretest performance rivals UNC’s posttest performance. Further instruction only
strengthened the effect. Even though other aspects of student understanding (particularly those associated with the meaning of spread) appear to be somewhat more difficult to learn, the experience with “Sizing up the book” demonstrates that small interventions directed specifically at student understanding of measurements can have big impact on student practice.

The findings of this study are similar to many findings in physics education research. Class gains on conceptual tests are often disappointing. Even the most effective, research-based introductory mechanics courses rarely raise the scores on conceptual tests (like the Force Concept Inventory) by more than 50% of the possible gain. (Hake 1999) That means that a class that scores 20% right on the pretest typically scores less than 60% right on the posttest, garnering less than half of the possible 80% gain on the test. On the positive side, studies in physics education research (Thornton and Sokoloff 1997; Abbott, Saul et al. 2000) have also shown that interventions as short as two hours can have tremendous impact on conceptual understanding. The present study provides further support of these findings.

5.3 Relationships among various modes of thinking

Prior to instruction, there is little indication of connections between various aspects of student performance. Four aspects of student understanding were assessed by multiple questions on the test designed for the present study: purpose of multiple measurements, meaning of spread, outlier treatment, and stacking. These constructs appear to have some validity. In addition to face validity, each construct appears to have some predictive validity. Gains in student understanding correlate well with observed differences in instruction. For example, SCALE-UP students, who received the most instruction in stacking showed the largest gain on items involving stacking. Further, correlations between different measures of student performance within each aspect are statistically significant, and, in some cases, quite strong. This suggests that the constructs developed for this study are internally consistent. Yet, there is little correlation between items assessing different aspects of student understanding. This suggests that the different aspects of thinking are unrelated.
In several studies, Allie and others (Allie, Buffler et al. 1998; Lubben, Campbell et al. 2001) uncovered a tendency among some students to place high importance on repeated values in a string of repeated measurements. On the basis of this and related observations, they devised the “point/set” scheme for categorizing student understanding. At first glance, it seems like “point/set” reasoning might underlie many student difficulties. For instance, one might suspect that “point/set” and ruler reading practice are related. If a student only reads the sure digits from a measurement device, the student will routinely encounter repeating values when performing multiple trials. Such students would be likely to display “point” reasoning, since a recurring value would indicate the “correct” value for the measurement. Students who interpolate are far less likely to encounter repeating values when performing repeated measurements. Interpolators would likely show signs of “set” reasoning, since they are likely to develop tools, like mean and uncertainty, for coping with the natural variation in results. A relationship between ruler reading practice and “point/set” reasoning, if it exists, was not discovered by this study. In fact, no correlation was found between “point/set” and any of the other three aspects of performance considered in this study.

Taken on its own, the failure of the “point/set” scheme to predict other aspects of student measurement behavior is unsatisfying, since the null result could be due either to the lack of an underlying relationship or the inability to accurately classify students using the scheme and questions published by Allie et al. As noted in previous chapters, student responses to the two questions used to assess “point/set” reasoning do not always contain elements of either “point” or “set” reasoning. However, the absence of a relationship between any of the aspects of thinking in this study, despite the internal consistency and face validity of the constructs, strongly suggests that there may not be a single underlying cause for the student difficulties observed in this study. Students rarely receive focused, coherent instruction on measurement practice and theory, so it is plausible that student “understanding” consists of a set of inconsistent, unrelated rules rather than a coherent (yet non-expert) framework of ideas.
5.4 Implications for instruction

This study shows that students can learn a considerable amount about measurement and uncertainty in one semester of lab instruction that accompanies a typical introductory physics course, provided the topics are addressed. Lab exercises in which students must calculate uncertainty and use the value to assess the agreement are reasonably effective at improving student understanding of the meaning of spread. Such activities need not be specifically directed at measurement concepts, and can be effective when incorporated into lab activities on various traditional physics topics. Lab instruction that forces students to choose methods of measurement and discuss the reasons for the choices seems to be effective at improving student measurement practices.

One of the principal difficulties students have is internalizing the purpose of the uncertainty concept. Most students seem to understand the purpose of constructing a mean after instruction, but very few students have a coherent picture of uncertainty and what the concept is for, even after instruction. Considerable effort needs to be put into helping students with this transition. Activities where students must assess the agreement of experimental results and justify their judgment appear to have some effect.

The SCALE-UP learning activities related to “stacking” are very effective. The single ten minute activity called “Sizing up the book” appears to be especially powerful. In the activity, called “Sizing up the book,” students estimate the thickness of a page in the textbook using a ruler. “Stacking” is an obvious approach to the problem. Without stacking, it is impossible to get even an order of magnitude estimate of the page’s thickness using a ruler. Students then use the estimate to “measure” the thickness of a period in a sentence in the text. This activity was done in the hour preceding the pretest in SCALE-UP. Not only did students perform well on a stacking question involving the thickness of floppy disks on the pretest, but nearly 40% of the SCALE-UP students also employed stacking to find the mass of a paper clip on the pretest. This indicates that many of the SCALE-UP students were able to transfer their understanding to a different context after a very short instructional intervention.

The lack of strong relationship between different aspects of reasoning suggests that students do not have a coherent understanding of measurement and uncertainty. The connection between measurement practices, data reduction practices and the underlying concept of uncertainty must be made more apparent to
students. Present texts fail to make many of the connections and students are often forced to rely on a collection of disconnected rules.

5.5 Implications and directions for future research

A model for student understanding of measurement and uncertainty with predictive power has yet to be developed. The “point/set” scheme of Allie and others as well as the “point/approximate/interval” scheme of Psillos et al. seem like good starting models for describing student thinking. Both schemes provide plausible explanations for a variety of student behaviors and suggest that there is an underlying cause for most student difficulties. However, the lack of any strong relationship between any of the aspects of performance examined in this study suggests that there may not be a single underlying cause for student difficulties observed in this study.

This study uncovered a weakness in the two questions that were used to classify students into the “point/set” scheme. Many student responses to the two questions published by Allie et al. fail to address the “point/set” dichotomy. New questions should be developed to improve the classification process. Used in study like the present one, improved questions would either uncover a connection between “point/set” reasoning and other modes of thinking or provide stronger evidence for the null result.

Very few students appreciate the need for the concept of spread and even fewer seem to be able to use the concept of uncertainty in a consistent manner. Instruction appears to help, but even after instruction, a vast majority of students continues to hold novice views about the nature of experimental data. Efforts need to be made to develop and evaluate activities and representations designed to appreciate the role of uncertainty in data collection and analysis. It should be noted that disappointing gains in conceptual scores are not uncommon in physics education. Outstanding research-based curricula rarely achieve more than 50 % of the possible improvement in scores on conceptual tests. Many alternative conceptions/frameworks persist in students despite carefully designed instruction.

Students in SCALE-UP learned to apply stacking in appropriate contexts. This result merits further investigation. Anecdotal evidence from the SCALE-UP classroom and elsewhere suggests that some students
hold strong reservations about stacking in some contexts, such as measuring the period of the pendulum. Despite this, SCALE-UP students employed the technique considerably more often than the other groups on both pretest and posttest. Observing students doing the “Sizing up the book” activity might explain why SCALE-UP students used stacking in such large numbers. The present study provides little insight into student thinking concerning stacking. Analysis of student conversations about stacking in a variety of measurement contexts might provide insight into student reasoning about stacking and about measurement in general.

Little has been done to assess the understanding of measurement and uncertainty in upper level undergraduates in the sciences. Even the uppermost levels of understanding described in present models of understanding are far from expert. Results from the present study suggest that many students view agreement between experimental results as binary: results agree or they don’t. Higher levels of novice understanding of the nature of experimental data can be mapped by interviewing upper level undergraduates and graduate students in a variety of fields.
Bibliography


Appendices
Appendix A: Lab instructions and assignments for students in SCALE-UP

In SCALE-UP, there was no lab manual. Lab instructions and assignments were delivered over the web. In most cases, instructions for the lab were posted on the class website. In other cases, the instructions were delivered using Webassign, a web-based homework system. Students did prelab and postlab assignments in WebAssign. This appendix includes the lab assignments and instructions for the SCALE-UP class as they were delivered in Spring 2002.

The materials are organized by the order in which the labs were performed. Each section opens with teacher notes. Materials delivered to the students follow.
A.1  Pennies (Lab 1)
Pennies Lab (Teacher notes)

Time: Plan for 30 to 45 minutes
Topic: Measurement, Process of Science

Overview:
In this lab, students formulate and test a hypothesis about the mass of pennies as a function of time. The results of this experiment are surprising to students. In 1982, the U.S. mint changed the composition of pennies. No other trend in the mass-time data is apparent. The lab has no physics overhead and serves as an introduction to error analysis and hypothesis formation and testing.

Equipment and Materials:
Entire Class:
• samples of pennies from each year from ~1970 to the present sorted by year.
• at least one triple beam balance per table
Notes on Setup:
• Samples should vary in size from 1 or 2 pennies to about 10 pennies, so that the uncertainty in the average mass of a sample varies from year to year.
• Divide the responsibility for data collection over the whole class and collect the communal data at the instructor’s station.
• Prepare a parallel activity for students to do while other students take measurements

Objective(s):
Measure mass using a balance.
Record measurements with estimated uncertainty (or correct number of significant figures)
Formulate and test a hypothesis.

Student Misconceptions:
Unknown

Other Student Difficulties:
Students may make untestable hypotheses by suggesting a cause (wear of coin, buildup of grime) for the mass trend over time.
Students may have different levels of understanding of error analysis and terms

Prerequisites:
None. Some level of error analysis is required for hypothesis testing, but student background can be augmented by appropriate supplementary materials.
Lab 1 (Pennies) Prelab

About this assignment

This must be submitted by each individual before attempting the lab.

1. [96621]

Info about Lab 1

In this lab you will investigate how the mass and thickness of pennies changes over time. In the process, you will practice lab skills that you will be using throughout the semester, including making a hypothesis, selecting a technique for taking high quality measurements, estimating the uncertainty in a single measurement and using Excel to create a graph.

A large sample of pennies has been collected and sorted by year of minting. Each table will get several samples of pennies to measure. Each sample will consist of several pennies from the same year. The average mass and average thickness of each sample will be measured. The data from the entire class will be collected and combined into one data set which each group will analyze.

Making a hypothesis

Before you do an experiment, it is important to take inventory of your own ideas about what you expect the data to show. This allows you to compare what the outcome of the
experiment with your expectations. A hypothesis is a **testable** prediction about the expected outcome of the experiment. What do you think the data will show? The hypothesis doesn't need to be correct, but it must be testable (so you can see if your initial thinking was on track). Write down **two** hypotheses for this experiment:

| Key: There should be one hypothesis about mass and one about thickness. The hypotheses do not have to be true, but they MUST be testable. Some examples follow. Testable hypothesis 1: The average thickness of pennies decreases with time. Testable hypothesis 2: The average mass of pennies decreases with time. |

It is also important to identify the reasoning behind your predictions. Why do you think the data will turn out the way you predicted?

| Key: There is no right or wrong answer, but it should be logical. For the hypotheses given above, an explanation might be that pennies lose copper as they get handled and abused. |

**Selecting a measurement technique**

Making good measurements is often simply a matter of choosing the best technique. Throughout the semester, you will be asked to the best measurements you can using the available equipment. In this lab, you will measure the average thickness of pennies from a given sample using a ruler. There are two ways to do this:

- **Method A:** Measure the thickness of each penny in the sample individually and take the average.

- **Method B:** Measure the thickness of a stack of all the pennies in the sample. Divide the thickness of the stack by the number of pennies in the stack to get the average.
Which method will give a more reliable measurement for the average thickness? Why is this method more reliable?

**Key:** Method B is better, since the uncertainty in the total mass gets "spread out" over the entire sample. Suppose the scale reads to the nearest 0.01g. If you measure 5 pennies, the total mass of all five is known to within 0.01. The average mass of a single penny is known to one fifth of 0.01g.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
Lab 1: Pennies

Background

In this lab you will investigate how the mass and thickness of pennies change over time. In the process, you will practice lab skills that you will be using throughout the semester, including making a hypothesis, selecting a technique for taking high quality measurements, estimating the uncertainty in a single measurement and using Excel to create a graph.

A large sample of pennies has been collected and sorted by year of minting. Each table will get several samples of pennies to measure. Each sample will consist of several pennies from the same year. The average mass and average thickness of each sample will be measured. The data from the entire class will be collected and combined into one data set which each group will analyze.

Taking the data

- Pick out a sample of pennies to measure. (One sample per group). Check the sample to see if all the pennies in the sample are from the same year. Put the misplaced pennies in the cup at your table marked "year?".
- Determine the average mass of a penny for the sample as accurately and precisely as you can. (Hint: Should you measure the mass of each penny in your sample or measure the mass of the entire sample? Why?) Determine the uncertainty in the average mass of a penny for each sample. The uncertainty is a numerical estimate of how reliable your measurement is. Is your value for average mass of the pennies in your sample within 1g? 0.1g? 0.01g? 0.00001g? Use your own judgement and common sense to obtain a numerical estimate how closely you determined the average mass of a penny. Record this experimental uncertainty (and the logic behind it).
- Determine the average thickness of pennies in your sample as accurately and precisely as you can. Determine the uncertainty in this measurement. (If your group has not done the mass measurement yet, look at those instructions for some helpful hints).
• Swap samples with a group at your table. Redo the measurements and compare your results with theirs. Did they do the measurement correctly? Did they estimate the uncertainty correctly? Did they do the rounding correctly?

• Once the data has been checked, select one person from your table to record your table's data in the class spreadsheet at the instructor's station.
  o Year of sample
  o Average mass of penny (in g)
  o Uncertainty in mass (in g)
  o Average thickness of penny (in mm)
  o Uncertainty in thickness (in mm)

Making sense of the data

• Make a graph of mass versus year using the data from the class spreadsheet. Include error bars for each data point. Save your finished graph. (Check out the tip on exporting data to find out how to use ftp: to save the file to your unity account).

• Does the graph support your hypothesis? Do the data suggest an alternative explanation?

• Which features on the graph could be attributed to the uncertainty in your measurements? Which ones can't be? Explain how you can tell.

Assignments (and other important grade info)

• Lab 1 Prelab WebAssignment that should have been submitted by each group member before you started the lab.

• Lab Report:
  o Each group of three students will submit a single written lab report via WebAssign.
  o The report will be due one week from today.
  o Very detailed information about what to put in the report (and how the report will be graded) are outlined in the grading rubric.
  o Information about uncertainty and sig fig's are available online from the Lab Info page
  o For this lab, you will only be graded on the Introduction, Results, Conclusions, Data, Presentation, and Overall Impact sections. For this particular lab (only) we would like you to
use this Microsoft Word document as a starting point. Feel free to modify it as you see fit.

- Teamsmanship Evaluation
  - Each individual will need to fill out a Teamsmanship evaluation. (The form will be available for submission via WebAssign.)
  - It is due one week from today.

- Post-Lab
  - Each individual will complete a Post-Lab activity on WebAssign.
  - This is also due one week from today.

Revised 1/02 by DSA
PostLab: Pennies

About this assignment

Each student does the postlab assignment. The postlabs focus on skills and ideas that were covered in the lab you just did.
Determine the thickness of a typical CD case as accurately and precisely as possible using the picture provided. Estimate the uncertainty in that measurement. In the answer box below, record the measurement and its uncertainty (using the proper number of significant figures). Note: The "thickness" is the dimension that the meterstick is set up to measure.

Thickness = [1.02] cm
Uncertainty in Thickness = [0.02] cm

*Explain* how you determined the uncertainty.
The answer depends on which picture was delivered. For the picture with the quarters, the uncertainty in the length of the stack is about 0.3mm (The ruler can be read to about 1/4 of the distance between the smallest divisions on the ruler). The uncertainty in the average mass of quarters is the 0.3mm divided by the number of quarters (17), or about 0.02mm (Notice that the uncertainty is rounded to one sig fig). For the picture of the CD's, the ruler is much harder to read. The uncertainty in the length of the stack is about 2mm. To get the best measurement, you should measure the length of 15 CD's (or so), so the uncertainty in the average thickness is 2mm divided by 15, or about 0.1mm.
A.2 Spring Lab (Lab 2)
Spring Lab (Teacher notes)

**Topic:** Process of science, measurement

**Time:** Plan for 60 to 80 minutes

**Overview:**
In this lab, students design a procedure to develop a mathematical relationship between the force applied to a spring and the amount it stretches. They then compare the model to Hooke’s law. We use springs that are non-linear below a certain load and linear above. Many real world springs (like screen door springs, carburetor springs) exhibit this behavior. This behavior appears to be common to springs that have no space between the coils at no load.

**Objective(s):**
- State hypotheses
- Design a procedure to determine the mathematical relationship between two variables
- Use Excel curve fitting tools to determine the mathematical relationship between two variables
- Compare theoretical model and real world data and develop explanations for differences
- Accurately determine the spring constant of the spring.

**Prerequisites:**
- Definition of hypothesis
- Some error analysis experience

**Equipment and Materials:**

Per group:
- table mount or spring
- Cenco conical spring* (see SHM lab)
- slotted masses and hanger
- meterstick
- computer w/ Data Studio, Excel, or some other program that includes fractional power fits

* Some Notes about Springs:
- Cenco conical springs are not cheap! (cost ~$20/spring)
- Springs deform with age and typically lose the low load non-linearity as they age.
- Several cheap springs (cost < 2$/spring) from Ferguson’s hardware also display this non-linearity at low load. Desirable springs for this lab have no space between the coils at zero load and have k<50N/m.
**Student Misconceptions:**

Students apply Hooke’s law blindly, dividing the applied force by the “amount of stretch” to find the spring constant. This method only works when the “amount of stretch” is also measured at zero force. It typically gives $k$ values that decrease with mass. A more robust experimental method is to graph the data and use the slope of the resulting line to find the spring constant.

**Other Student Difficulties:**

Students may not take enough data over a wide enough range of applied forces. In particular, students may not adequately explore the low load region. Students misuse fit tools, often employing 3rd and 4th order polynomials to fit the entire data set. Student problems with operational definitions of measurement cause results that vary from group to group and make comparison and/or reproduction of results tricky, if not impossible. For instance:

- Many students will not record how they measure the “amount of stretch.”
- Students may not measure the length of the spring by measuring the distance between two points on the spring. (For instance, some groups included the length of the weight hanger in the length of spring measurement).
- Some students will define zero load as no weight on the weight hanger (which itself has a mass of 50 grams). Groups that do this may miss the non-linearity of the graph altogether.
Prelab for Lab 2 (Springs)

About this assignment

This must be submitted by each individual before attempting the lab.

1. [97873] In the lab you will be doing Wednesday, you will be using Excel's fitting tools. Read through the Tips on Fitting from the class website. Once you've read the tips, examine the two graphs below (the data should look very familiar!). The two graphs are identical except for the fit equation. Which fit equation is better? Why?

Key: The simpler fit (fit #2) makes much more sense. It strongly suggests that pennies before 1982 were all about the same mass, while pennies manufactured after that date had a higher (but constant) mass. The polynomial fit to the data suggests that there is some complicated relationship that explains how the mass is related to the year of minting.
Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
Spring Lab

Background: Developing mathematical models

Scientists and engineers often perform experiments to determine how one variable affects another. Using graphs and fitting tools, scientists and engineers can produce simple mathematical models which predict how a physical system works. In this lab, you will produce a mathematical model for springs. The data you take will answer two basic questions:

- What is a simple mathematical formula that relates the amount a spring stretches to the weight hung from it? How general is this relationship?
- How do the numbers in a experimental determined formula relate to physical properties of the spring?

Equipment

- Spring
- Table clamp
- Mass hanger (50 g) and standard masses
- Meterstick

Precautions

- Do not stretch the spring beyond its elastic limit! Hanging too much weight from the spring or pulling it too hard may ruin this expensive spring.

Taking and recording data

1. Before you take data, plan a procedure for taking and recording data. Keep in mind your goal is to determine the mathematical relationship between the force applied to the spring and the amount it stretches. Some questions you should consider are:
   - How many data points should you take?
   - Which data points should you take?
   - How will you measure the length of the spring (from where to where)?
   - Should you measure the weight of the hanger? the weight of the spring?
2. Execute your procedure. As you do this,
1. **Take careful notes!** Your notes should be detailed enough that someone could reproduce your data from just your notes. Two things you should include: are a diagram of the experimental apparatus that shows how you got your measurements and an estimate of the uncertainty in your measurements (and a brief description of how you estimated the uncertainty)

2. **Graph your data in Excel as you take it.** Seeing the graph may help you make choices about what data to take. Don't worry if you end up taking more data (or different) data points than you had originally planned. Be attentive to surprises and follow up on hunches!

3. **Don't dismantle your apparatus (or leave class) before you analyze and interpret your data.**
   - **Check your data:** Does your data cover the whole range? Do the data show a distinct pattern? Are there data points that don't fit your pattern? Keep in mind that you are graded on the quality of the data and the accuracy of your conclusions!
   - Determine the equation(s) that relates the amount the spring stretches to the force applied to the spring. Check out the tip on how to use fitting to produce meaningful results.
   - Check to make sure your equation gives accurate results. Your equation should make predictions for the stretch that are within the estimated uncertainty.
   - Interpret the numbers in the fit equation you get. How are the numbers the equation related to physical properties of the spring? [Hint: If you're not sure what the numbers mean, figure out what units the numbers have.]
   - Compare your results to other groups at your table. Are the shapes of the graphs the same? Are the equations the same? Are the numbers in the equation the same? Should they be?
PostLab 2 (Springs)

About this assignment

Each student does the postlab assignment. The postlabs focus on skills and ideas that were covered in the lab you just did.
1. A company that makes screen doors is testing springs for use as a return spring for their newest model. To test the stiffness of the spring, engineers record the length of the spring when various amounts of mass are hung from the spring. Data for one of the springs are below.

<table>
<thead>
<tr>
<th>Mass (g)</th>
<th>Spring Length (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.7</td>
</tr>
<tr>
<td>50</td>
<td>7.7</td>
</tr>
<tr>
<td>100</td>
<td>7.7</td>
</tr>
<tr>
<td>150</td>
<td>7.7</td>
</tr>
<tr>
<td>200</td>
<td>8.1</td>
</tr>
<tr>
<td>250</td>
<td>8.4</td>
</tr>
<tr>
<td>300</td>
<td>8.9</td>
</tr>
<tr>
<td>350</td>
<td>9.3</td>
</tr>
<tr>
<td>400</td>
<td>9.8</td>
</tr>
<tr>
<td>450</td>
<td>10.2</td>
</tr>
<tr>
<td>550</td>
<td>11.1</td>
</tr>
<tr>
<td>650</td>
<td>11.9</td>
</tr>
<tr>
<td>750</td>
<td>12.8</td>
</tr>
<tr>
<td>850</td>
<td>13.7</td>
</tr>
<tr>
<td>950</td>
<td>14.7</td>
</tr>
</tbody>
</table>

Use the data above to make a graph of weight vs. spring length and then answer the questions below. (Note: If you are using Internet Explorer, you can select the table and paste it directly into Excel).

The engineers want to make a mathematical model for the spring. What type of equation should they use to fit the data? Should all of the data be included in the fit? Explain your answers.

Key: Above about 200g, the data look linear. Use a linear fit equation, but exclude the first four points.
Use a trendline to determine the spring constant, \( k \), for the spring. (\( k \) is number, usually measured in N/m, that describes the stiffness of the spring).

\[
k = [1.116] \text{ N/cm}
\]

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
A.3 Friction (Lab 3)
Friction Lab

**Topic:** Process of science, measurement  
**Time:** Plan for 40 to 60 minutes

**Overview:**

First, students design a procedure to measure the coefficients of static and kinetic friction between the force probe housing and a sheet of Xerox paper. Students then design a procedure to test whether the coefficient of kinetic friction depends on the normal force or the area of contact surface.

**Objective(s):**

- Apply Newton’s 2nd law and the friction model to measure coefficients of friction
- State hypotheses
- Design a procedure to determine the mathematical relationship between two variables
- Estimate the uncertainty in a calculated quantity.
- Evaluate hypotheses using results and error estimates

**Prerequisites:**

- Definition of hypothesis
- Friction Tangible
- Some error analysis experience

**Equipment and Materials:**

Per group:

- ULI or PASCO force probes with appropriate software
- Weight set
- Sheet of Xerox paper, tape
- Rubber bands (to pull the probes with)
- Spreadsheet with graphing and linear fitting

**Student Misconceptions:**

Many students confuse force of friction and coefficient of friction.

**Other Student Difficulties:**

Students may not take enough data over a wide enough range of normal forces. Many students will try to record all F vs. t and may not label their data files sufficiently. Student reports include too much meaningless F vs. t data, rather than report only the plateau values.
PreLab 3 (Friction)

About this assignment

This must be submitted by each individual before attempting the lab.

1. [98264] Consider the following two situations:

   **Situation A:** A force probe is pulled across a sheet of paper at constant velocity.
   **Situation B:** A force probe with a 500 gram mass on top is pulled across a sheet of paper at constant velocity.

In which situation is the frictional force larger? Explain.

Key: The frictional force is greater in Situation B. The force of friction increases as the normal force increases.
In which situation is the coefficient of kinetic friction larger? Explain.

Key: The coefficient is the same for both, because the coefficient of friction depends on the type of surfaces that are sliding. Both the friction force and the normal force increased, leaving the coefficient of friction unchanged.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
Friction Lab

Equipment

- Force Probe(s)
- Lab interface
- Computer with data acquisition software
- Sheet of laser printer paper

Your tasks

1. Set up the interface box and force probe. (Click here for instructions on setting up the equipment).
2. Determine the coefficients of static and kinetic friction between the plastic of the probe housing and the sheet of paper provided (DO NOT USE NOTEBOOK PAPER).
3. Determine whether these coefficients depend on the force pushing the surfaces together.
4. Determine whether these coefficients depend on the area of surface contact.

Hints for taking data

- For task 2, make sure you run a few trials, using the same setup each time. This will allow you to estimate how reliable your measurements are.
- For task 3, you should take enough data to produce a meaningful graph.

Some Lab Report Suggestions

- *Meet with your entire group as soon as possible to do the calculations.* Make sure each person understands how to get the results from the data and that each person understands how the results relate to both hypotheses.
- *Finish the first draft a couple of days before the lab is due* and *have a friend grade the report*, using the rubric as a guide.
PostLab 3 (Friction)

About this assignment

Each student does the postlab assignment. The postlabs focus on skills and ideas that were covered in the lab you just did.

1. A group of students are writing up the friction lab. During the first part of the experiment they had done five identical trials, in which the probe was pulled across a piece of paper. They recorded the maximum force they could apply before the probe started to slide. They also recorded the force applied to drag the probe at constant velocity. Their data for this part of the lab are in the table below.

Weight (probe only) = 3.31 N

<table>
<thead>
<tr>
<th>$f_s$ (in N)</th>
<th>$f_k$ (in N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.77</td>
<td>0.91</td>
</tr>
<tr>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td>0.68</td>
<td>0.88</td>
</tr>
<tr>
<td>0.74</td>
<td>1.10</td>
</tr>
<tr>
<td>0.73</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Use the group's data to find the static coefficient of friction and its uncertainty. Report the value below as it should appear in their lab report. Make sure that you report the correct number of significant figures in your answer.
Briefly explain how many sig fig's each number should have and why.

Key: There are several acceptable methods for estimating the uncertainty, which yield different values: The simplest (but least correct) is to take the difference between the highest and lowest value for \( \mu \) and divide by two. More correct is to find the average deviation. Even better is to calculate the standard deviation. Whichever method is used, the uncertainty lies between 0.02 and 0.05 and should be reported to only one sig fig. The value for the coefficient should be rounded off to the hundredths place so that the digits match the uncertainty.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
A.4 Rolling Objects (Lab 4)
Rolling Objects Lab (Teacher notes)

**Topic:** Process of science, rolling motion, energy conservation  
**Time:** Plan for 40 to 60 minutes

**Overview:**  
Students roll a variety of objects down an incline to determine which characteristics of the object affect the time it takes to roll down the incline.

**Objective(s):**
- Design and execute a controlled experiment
- Assess agreement of two experimental results using uncertainty
- Estimate the uncertainty in a measured quantity using multiple trials

**Prerequisites:**
- Definition of hypothesis  
- Some uncertainty analysis experience

**Equipment and Materials:**

Per group:
- 1.2 meter Pasco tracks
- Rolling object set (see comments below)
- stopwatches
- spreadsheet with graphing and linear fitting

**Comments:**
- The set of rolling objects should have at least pairs of objects that differ by only one attribute. At NCSU, a special set of cylinders was fashioned by the machine shop from aluminum and steel. (see RollingObjectsList.htm). In 2002, we purchased moment of inertia rings with movable masses (so that mass distribution can be changed while mass and radius are held constant).
- Include a PASCO cart in the set of objects. Students should use this as a calibrator for ramp angle. (Measuring ramp angle is often unreliable, since tables are rarely level).

**Student Misconceptions:**
- Students may think mass and radius matter.

**Other Student Difficulties:**
- Students may not be able to pick pairs of objects for meaningful comparisons. Comparisons must be made in a logical order. Once one factor (e.g. mass) has been ruled out as factor in the time it takes, comparisons can be made with unequal masses. (This assumes that the two variables don’t interact).
- Students may attempt to compare data without accounting for differences in the angle of incline.
Quiz 6: Rolling Objects

About this assignment

This is a group activity to be done in class. Each person needs to submit this assignment.

Enter the password for this assignment here: [98943]

1. [98943]

In Monday's lab, you will investigate which characteristic(s) affect the amount of time it takes an object to roll down an incline. Taking precise and accurate time measurements is essential for this lab. In this prelab activity, you will learn how to take accurate data for Monday's lab.

Do not take data yet! Since you will be using the same incline throughout Monday's experiment, setup is critical. How can you get the most precise time measurements? Should you use a steep incline or a shallow one? Should you time the object over a short distance or a long distance? Explain the reasoning behind your choices.
Key: In order to get the most precise time measurement, each run should last as long as possible. That way, the percent uncertainty will be as small as possible. A long distance and a shallow incline will work best. The incline should be steep enough that the object rolls easily, though.

Set up the incline. Record the setup info here:

Incline angle = \( [4]^\circ \)  
Distance (that the objects will travel) = \( [95] \) cm

Note: Webassign will accept any number of sig figs for your answers. However, you will need to pay attention to sig figs in Monday's lab.

Now you are ready to take data. For this prelab, you will time the dynamics cart and one other object. Select one other object to time from the list below:

Poll questions: Please give a candid response to survey questions. There is no incorrect response (blank responses do not count).

Time how long it takes each object to travel down the incline. Do five trials for each. Record your data here:

<table>
<thead>
<tr>
<th>Trial</th>
<th>Time (s)</th>
<th>Dynamics Cart</th>
<th>Other Object</th>
</tr>
</thead>
</table>
Report the results below:

<table>
<thead>
<tr>
<th></th>
<th>Average Time (s)</th>
<th>Uncertainty (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamics Cart</td>
<td>[\text{[0]}]</td>
<td>[\text{[0.01]}]</td>
</tr>
<tr>
<td>Other Object</td>
<td>[\text{[0]}]</td>
<td>[\text{[0.01]}]</td>
</tr>
</tbody>
</table>

For info about estimating uncertainty for multiple measurements, consult this page.

According to your data, which object takes less time to get down the ramp?

Key: The dynamics cart is faster.

Is it likely that the difference in times between the faster object and the slower object is due to uncertainty in measuring the time? Explain your answer carefully, using the concept of standard error.

Key: The difference in times is much larger than the standard error. It is very unlikely that the difference in times is due to random variations in timing.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
PreLab 4 (Rolling Objects)

About this assignment

This must be submitted by each individual before attempting the lab.

1. A dynamics cart (m=339 g) is released from rest and rolls without slipping down a 1.0 ° incline which is 100.0 cm long.
   a) Use the principle of energy conservation to predict the speed of the cart at the bottom of the ramp.  \( \sqrt{0.585} \) m/s
   b) Predict the time it takes the dynamics cart to reach the bottom of the ramp.  \( \sqrt{3.42} \) s

Suppose the measurement is repeated with two different objects from this list.

   c) Predict the time that it takes object 2 to reach the bottom of the ramp.  \( \sqrt{4.84} \) s
   d) Predict the time that it takes object 3 to reach the bottom of the ramp.  \( \sqrt{4.84} \)

2. Some of the objects in this list will travel down an incline in less time than others. Rank the objects according to how long each will take to roll a fixed distance down an incline. Explain the reasoning behind your prediction. (Assume each is released from
The only thing that affects the time is the ratio of $\frac{I_{cm}}{MR^2}$. The hoops have a higher percentage of their KE tied up in rotation (and therefore less KE left for translation). (see Serway and Beichner, p.330 for a complete explanation).

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
About this assignment

Complete this group assignment before taking data for the lab.

1. [99028]
   Acceleration of Rolling

   Objects

   Introduction:

   What characteristic(s) of an object affect the amount of time it takes to roll down an incline? Which characteristics don't? In this lab you will design a procedure to answer these questions.

   Equipment available:

   1.2 m track, various cylindrical objects (click here for details), fixed and movable mass
Planning for Data Taking

1. The goal of this lab is to find out which factors affect the time it takes an object to roll down the ramp (and which factors don’t). Ideally, as an experimenter, you’d like to keep all variables (but one) constant, so that you can determine whether that factor influences the dependent variable. Sometimes (like in this lab), it is difficult (or impossible) to vary only one thing at a time.

2. In your group, brainstorm two testable hypotheses. State the two hypotheses here. Label them H1 and H2.

   **Key:** Two examples of testable hypotheses include: (H1) the cross-sectional shape (whether the mass is near the edge or near the center) will affect the time, (H2) length of the object will not affect the time.

3. Identify at least three objects you will time to test H1:

   **Key:** Objects to use to test h1: Fixed Mass Ring, Movable Mass Ring (time it with the mass in the middle and then with the mass at the edge.)
4. Identify the key similarities/differences between the objects you will compare.

Key: Similarities/differences for objects use to test h1: All of these have the same mass, radius, length, but vary in how the mass is distributed.

5. Explain how you would use the data to show that H1 was supported (or refuted):

Key: How the data would have to come out to support (or refute) H1 given above: Difference in times between the various objects will be larger than the uncertainty in the times if h1 is true.

6. Explain any difficulty you might encounter in making a conclusion about H1 from your data.

Key: Difficulties interpreting data for the H1 given above: If the uncertainty in time is small enough, there should be little trouble interpreting the data. The only thing being varied is the variable of interest.
7. Identify at least three objects you will time to test H2:

Key: Objects to use to test the H2 given above: Long PVC, short PVC with same radius; Long Al rod, short Al rod with same radius.

8. Identify the key similarites/differences between the objects you will compare.

Key: Similarities/differences for objects use to test the H2 given above: Each pair has same radius and composition. They differ in length and also in mass.

9. Explain how you would use the data to show that H2 was supported (or refuted):

Key: How the data would have to come out to support (or refute) the H2 given above: IF there is no difference in time for either pair, this supports H2.

10. Explain any difficulty you might encounter in making a conclusion about H2 from your data.
Key: Difficulties interpreting data for the H2 given above: If the times are about the same for each pair, it MIGHT be because the effect of increasing the length and increasing the mass cancel each other out. The data cannot rule out this possibility.

11. Coordinate with the other groups at your table. Choose the three best hypotheses from your table to test and report them here:

Key: The table’s best hypotheses.

12. Briefly explain why these are the three best hypotheses.

Key: Why they are the best.

Taking Data:

Assign one group at your table to test your table’s best three hypotheses. Take data. Use the best possible setup and measurement techniques you can! High quality data is essential in this lab.
Collect the data from all the groups at your table. You will be expected to analyze the entire table’s data in your report.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
Postlab 4 (Rolling Objects)

About this assignment

Each student does the postlab assignment. The postlabs focus on skills and ideas that were covered in the lab you just did.

1. A student times how long it takes a solid aluminum rod to roll 100.0 cm down an inclined plane. The student uses a stopwatch which can be read to the nearest 0.01 s. She is careful to release the rod from rest each time. The times recorded by the student are 4.39 s, 4.32 s, 4.45 s, 4.31 s, and 4.45 s. Compute the average time and the uncertainty in the time for the aluminum rod. Report the values below.

   Average time = \[4.38\] (significant figures = 3)
   Uncertainty in time = \[0.03\] (significant figures = 1)

The student then performs the experiment again. This time she uses a solid steel rod with the same dimensions as the aluminum rod she just used. The results are 4.03 s, 4.04 s, 4.15 s, 3.98 s, and 4.07 s.

   Average time = \[4.05\] (significant figures = 3)
   Uncertainty in time = \[0.03\] (significant figures = 1)
Can the student legitimately conclude that the steel rod was faster using this data? Explain, using numbers to support your answer.

Key: The student can claim the steel rod is faster. The two time ranges do not overlap. The time for the steel rod plus the uncertainty in the time for the steel is less than time for the aluminum minus the uncertainty in the aluminum's time.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
A.5 Pendulum (Lab 5)
**Pendulum Lab**  
**Type:** LAB  
**Topic:** Process of science, measurement  
**Time:** Plan for 60 to 80 minutes

**Overview:**
In this lab, students make and test hypotheses about the factors that influence the period of a pendulum. This is not an exposure-verification lab.

**Objective(s):**
- State hypotheses
- Design a procedure to test multiple hypotheses involving several independent variables
- Design a procedure to determine the mathematical relationship between two variables
- Use Excel curve fitting tools to determine the mathematical relationship between two variables
- Use uncertainty estimates to refute or support hypotheses
- Use graphical methods to determine the mathematical relationship between two variables

**Prerequisites:**
Definition of hypothesis  
Some error analysis experience

**Equipment and Materials:**
- Per group:
  - string
  - support for pendulum (stand, upright, arm, two clamps)
  - slotted masses and hanger
  - stopwatch
  - protractor
  - meterstick
  - computer w/ Data Studio, Excel, or some other program that includes fractional power fits

**Notes on Setup:**
- Slotted mass and hanger arrangement is an extended mass (on purpose). Students who have had physics will not know the results of the lab.
Student Misconceptions:

Some students may not yet understand that, in order to test the effect of a single factor, all other factors must be held constant. Some students believe that the period of a pendulum depends (strongly) on amplitude. Many students will not know how to measure the length of the pendulum. It is not important that the y measure to the center of mass, but they should record how they determine the length.

Other Student Difficulties:

Students do not take enough data! In Fall 2000 and Spring 2001, about 80% of the groups took three data points on each variable (including length). Most groups do multiple trials for the period, but in Spring 2001 measured one period (rather than taking the time or several periods and dividing by the number of swings). Students present meaningless graphs in their lab reports. Common mistakes include:

- making bar graphs
- putting all data points on the same graph (rather than making separate graphs for each variable)
- meaningless curve fitting (typically 2nd through 5th order polynomials for 3 point graphs)
- missing or inappropriate error bars

Students do not record the details of their measurement methods (e.g. a diagram for the length of the pendulum, whether multiple swings are used).
Prelab: Pendulum

About this assignment

This is a group activity to be done in class.
Enter the password for this assignment here: 

1. [99260]

Introduction

The period of a pendulum is defined as the amount of time it takes the pendulum to make one complete swing. In the pendulum lab, you will investigate which characteristics of a pendulum affect the period of the pendulum (and which ones don’t). In this assignment, you will plan your experiment.

Planning the experiment

What characteristics of the pendulum do you think will influence its period? Which characteristics will have no effect on the period?
Key: Some factors that students might identify include: length of the pendulum, mass of the bob, angle of release. Factors should be measurable.

Hypothesize about one of these factors in the space below. State your hypothesis in the form: Increasing the .... will increase/decrease/not affect the period.

Key: Student predictions do not need to be factually correct. Example: Increasing the mass will increase the period.

Design an experiment that will allow you to see if, and in what way, the factor you’ve identified affects the period of the pendulum. Describe in considerable detail what you would do to find out if your predicted idea is valid.

Key: Correct answers will vary. Some important details to include follow. The procedure should vary the characteristic being tested over as wide a range as possible while all other variables are kept constant. The procedure should recommend multiple trials for each measurement of the period. The procedure should describe how each of the variables is to be measured.

Getting High Quality Data

Taking precise and accurate measurements for the period is critical for this lab. Even though you will be using a stopwatch for this experiment, you can get high quality data by using good techniques. Consider the following two methods for measuring the period of a pendulum:

Method A: Release the pendulum from a small angle with the vertical. Use the stopwatch to time one full swing of the pendulum. Repeat four times (for a total of five trials), making sure the angle of release is the same each time. Report the average and standard error.
**Method B:** Release the pendulum from a small angle with the vertical. Use the stopwatch to time ten full swings of the pendulum. Divide the total time by ten to get the period. Repeat four times (for a total of five trials), making sure the angle of release is the same each time. Report the average and the standard error.

What advantage does Method B have over Method A? Explain.

**Key:** Method B will produce less variable results for the period. The natural variation in the total time will get divided by ten.

What is a potential drawback to Method B? Explain.

**Key:** Method B assumes that the period does not change as the motion dies out. If the last swing takes more time than the first, then Method B will give a result for the period which is too high.

Explain how could you test whether Method B gives correct results.

**Key:** One way to test is as follows: Time 10 swings five times, making sure the pendulum is released from the same initial angle each time. Compute the period and the uncertainty in the period. Then time 20 swings five times, keeping the same initial angle as before. Calculate this period and its uncertainty. If the two are the same (within uncertainty), Method B gives correct results (and should be used).

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
Pendulum Lab

Background

A pendulum is a mass that swings back and forth on the end of a string or rigid arm. Pendulums are used in mechanical clocks because they swing back and forth at regular intervals. In this lab, you will determine which factors influence the period of a pendulum and which ones don't. (The period is the amount of time it takes to go back and forth once.)

Instructions

Take high quality measurements of the period. The available equipment is capable of measuring the period with an uncertainty of less than 1% of the measured period. All of your measurements for the period should be at least that precise. Decide on a method for measuring the period. Test your method for measuring the period and make sure that your method meets or exceeds this level of quality.

Coordinate with the other groups at your table*. Each group at the table will thoroughly investigate a different factor that might influence the period of the pendulum. Take enough high quality data to produce a graph (7-10 points over a wide range) which shows how (not just whether) the factor you investigate affects the period.

Carefully record how you measure each of the characteristics of the pendulum in your notes. (For example, when you measure the length of the pendulum, where do measure from and to?)

Make sure you get (and check) data from the other two groups at your table. You will be expected to analyze the data from the other two groups (as well as your own data) in your report. Your skeptic should check all three sets of data. Do not leave class until the other two groups at your table are satisfied with your data.

* If there are fewer than three teams at your table, coordinate with two groups at another table.

Analysis:

- Construct three graphs (1 for each group’s data).
- Compare your group’s data to the group’s hypothesis.
• Compare the data from all three groups to what the textbook formula for a pendulum predicts. Explain any discrepancies.
• Use the data to find the local value for g.

Due Dates, etc.:

Full Report, Teamsmanship Evaluation, PostLab due in one week

Revised 4/02 by DSA
Postlab 5 (Pendulum)

About this assignment

Each student does the postlab assignment. The postlabs focus on skills and ideas that were covered in the lab you just did.

1. The animation below simulates the motion of a mass attached to a spring sliding across a surface. Click here to initialize the animation and the "Play animation" button to see it move. You may click-drag the blue block to change its position. You may also change the values of different quantities in the table to the right of the animation and then apply them to the animation with the "Apply values" button. Which of the different quantities that you can alter affect the period of the motion?

Write one hypothesis for this "experiment." (Make sure your hypothesis is testable.)
The hypothesis should be a testable statement about the effect of one variable on the period of oscillation.

Test your hypothesis and write a short, but convincing, argument either supporting or refuting your hypothesis. (The best arguments should contain a short description of your method, data from the animation and uncertainty estimates for your data).

Key: Answers will vary. Exemplar: I varied mass while keeping all the other factors constant and found that the mass affects the period. The difference between the period for m=1 (period quoted here) and m=100 (period quoted here) was much larger than the uncertainty in the period (number quoted here). I measured the time for 5 periods three times and got times that differed by about 0.1 s, making an uncertainty of 0.02 s.

Which of the following variables affect the period of the oscillating mass?

- [ ] Initial velocity
- [x] Spring strength
- [ ] Initial position
- [x] Mass of block

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
A.6 Simple harmonic motion (Lab 6)
Oscillations Lab (Teacher notes)

Topic: Process of science, measurement, simple harmonic motion, Hooke’s law
Time: Plan for 60 to 80 minutes

Overview:
In this lab, students will measure the spring constant of a helical spring using using Hooke’s law. From this, they predict the period of oscillation for two masses (100g and 300g). A massive spring (m=150g) is deliberately used so that the standard textbook formula does not give the right value for the period. Students use Interactive Physics to refine the model of the system and explain the observed discrepancy.

Objective(s):
- develop and execute a procedure to determine the spring constant of a helical spring using Hooke’s law (they must apply Newton II)
- develop and execute procedure to determine the spring constant of a helical spring using the period of oscillation
- (optional) use InterActive Physics to develop an improved model for the spring/mass system
- construct graphs, determine trendline equations using Excel and interpret the meaning of fit parameters (like slope and intercepts)

Prerequisites:
This is a hard lab! Students need knowledge of Hooke’s law, formula for period of spring/mass system. Facility with graph construction, Excel and graphical analysis is essential.
Familiarity with InteractivePhysics is required, if the IP part of the lab is used. Experience in designing lab procedures is also strongly recommended (see comments “Other student difficulties”).
Equipment and Materials:

Per group:

- CENCO conical spring
- Hardware for suspending spring (see notes below)
- Stopclock
- Masses
- Computer with InteractivePhysics
- Meterstick
- Triple beam balance

Notes on Setup:

- Warn students about elastic limit of spring
- Label the springs and have students record which spring they use
- Suspend spring narrow end up
- Physical constraints may limit the range of possible masses (mass touches floor at equilibrium). Try to use mounting equipment that allows about 1.5 m of spring extension

Student Misconceptions:

Some students may not yet understand that, in order to test the effect of a single factor, all other factors must be held constant.
Some students believe that the period of a pendulum depends (strongly) on amplitude.
Many students will not know how to measure the length of the pendulum. It is not important that the y measure to the center of mass, but they should record how they determine the length.

Other Student Difficulties:

It is possible that students might overstretch these expensive springs, but warnings in the web page have been sufficient so far…
Some students may hang the spring upside-down (narrow end should be at the top). If the clearance below the mounting position for the spring, students may not be able to use a wide range of masses.
If students are not trained to take sufficient data for graphical analysis, students may take only a few points or data that covers only a small range of masses
SHM InClass

About this assignment

In class assignment for Simple Harmonic Motion lab. Each group will input data and numbers calculated from the data, and the IP file from the last part of the lab. WebAssign checks to see if the data are reasonable, and whether the calculations are correct. [This is a group assignment]. There is no formal write-up for this lab. Your lab grade will be based on this in-class assignment.

1. [99359] Spring Constant (from Hooke's Law data): What is the value for the spring constant you found using Hooke's law?

   [8.5] N/m ± [2] N/m

   (Click here for a spreadsheet to calculate slope and its uncertainty.)

   Note: Webassign can only check if the values you input are reasonable. Webassign cannot check whether you did this part correctly.

   Briefly explain how you calculated this k value. Indicate which points you included in the trendline fit and state why you chose those points.
2. [61381] **Predicted Period (from textbook formula):** Use the measured value of the spring constant to predict the period of oscillation for a 100 g mass hanging from your spring. Calculate the uncertainty in the predicted period. DO NOT round off your answer to the right number of sig figs for this question.

\[ T_{\text{formula}} = 1.99 \pm 0 \]

**Note:** Webassign will check whether you did the calculation correctly, based on the value of \( k \) and its uncertainty that you gave in the previous question. It assumes you used the upper bound/lower bound method to calculate the uncertainty in the predicted period.

3. [61382] **Measured Period:** Measure the period of the real world spring/mass system. Estimate the uncertainty in the measured period.

\[ [0.8] \pm [0.0151] \]

**Note:** Webassign can only check whether the values you input are reasonable. Webassign cannot check whether you did this part correctly.

Explain how you calculated the uncertainty in the period.

---

**Key:** Students should measure the time for a fixed number of oscillations (like 5 or 10) and divide the time by the number of oscillations to find the period. They
should do this several times and take the standard error to find the uncertainty in the period.

Does the measured period of the real world spring/mass system match the period you predicted using the textbook formula? Explain, using numbers. (Make sure the numbers are rounded off to the right number of sig figs).

Key: The two periods will not be consistent (i.e. the possible ranges for the two values will not overlap).

Does the period of the Interactive Physics spring/mass system match the period you predicted using the textbook formula? What does this suggest about the model underlying the calculations Interactive Physics makes?

Key: The IP period matches the period predicted by the textbook formula, so the model IP uses is the same as the book.

4. Use Interactive Physics to create a spring/mass system which agrees with all of the data. Don't forget to record the main features of the spring/mass system in your notes. Upload the Interactive Physics file you created.

Key: IP file

To upload the file: Click 'Browse' and locate your file using the dialog box. Click on the file to select it and click the 'Open' button. The file will be uploaded when you submit.

Submit Sample for Grading

Note for faculty: numbers in the boxes above are exact. Students' keys will be shown rounded to three significant figures.
Appendix B: Uncertainty guide for students in SCALE-UP

This appendix contains the measurement and uncertainty guide that was posted on the SCALE-UP class website in Spring 2002. The first six sections of this guide borrow heavily from a similar guide written by Duane Deardorff (see the Bibliography of this dissertation). It should be noted that some of the terminology and presentation do not conform to the guidelines published by the International Standards Organization.
Measurement Uncertainty Guide

Contents

1. Introduction
2. Reporting experimental numbers
3. Comparing experimental numbers
4. Accuracy and precision
5. Random and systematic error
6. Sources of error
7. Estimating uncertainty from a single measurement
8. Estimating uncertainty from a repeated measurement
9. Estimating uncertainty in a calculated quantity
10. Common Mistakes

Adapted from D. Deardorff's Measurement and Uncertainty Guide by DSA
Revised 01/06/02
Introduction to Error Analysis

What is error analysis?

Some numerical statements are exact: Mary has 3 brothers, $2 + 2 = 4, \pi = 3.1415926...$ Others are not. Try measuring the diameter of a quarter. No matter what technique or tools you use, it is impossible to say exactly what the diameter is equal to. When you measure the same quantity by different methods, or even when making multiple measurements using the same method, you may obtain slightly different results. In other words, all measurements have some degree of uncertainty. Error analysis (also called uncertainty analysis) is the process of providing a numerical estimate of how "good" a measurement is.

Why do error analysis?

Error analysis is extremely important for interpreting the results of experiments. Suppose you want to find out whether the mass of an object affects how long it takes the object takes to fall to the ground. You measure the time it takes a brick to fall a certain distance to be 0.98 s and the time it takes a marble to fall the same distance to be 0.86 s. Which object took less time to fall? The answer depends on how reliable the time measurements are. Are these times measured reliably to the nearest 0.01 second or to the nearest 0.1 second? Without error analysis, it is impossible to say whether the difference in these two times is real or is due to inexact measurements.

Error analysis is inexact

Uncertainty analysis is the art of estimating how off we think we could be in our experiment. Many students lose sight of this. After all, if we knew exactly how much we were off by, we would know the actual value of the measurement (or result)!
Reporting Measurements and Experimental Results

Best Estimate ± Uncertainty

When scientists make a measurement or calculate some quantity from their data, they generally assume that some exact or *true value* exists based on how they define what is being measured (or calculated). Scientists reporting their results usually specify a range of values that they expect this "true value" to fall within. The most common way to show the range of values is:

\[
\text{measurement} = \text{best estimate} \pm \text{uncertainty}
\]

**Example:** a measurement of 5.07 g ± 0.02 g means that the experimenter is confident that the actual value for the quantity being measured lies between 5.05 g and 5.09 g. The uncertainty is the experimenter's best *estimate* of how far an experimental quantity might be from the "true value." (The art of estimating this uncertainty is what error analysis is all about).

---

How many digits should be kept?

*Experimental uncertainties should be rounded to one significant figure.* Experimental uncertainties are, by nature, inexact. Uncertainties are almost always quoted to one significant digit (example: ±0.05 s). If the uncertainty starts with a one, some scientists quote the uncertainty to two significant digits (example: ±0.0012 kg).

**Wrong:** 52.3 cm ± 4.1 cm  
**Correct:** 52 cm ± 4 cm

*Always round the experimental measurement or result to the same decimal place as the uncertainty.* It would be confusing (and perhaps dishonest) to suggest that you knew the digit in the hundredths (or thousandths) place when you admit that you unsure of the tenths place.

**Wrong:** 1.237 s ± 0.1 s  
**Correct:** 1.2 s ± 0.1 s
Comparing experimentally determined numbers

Uncertainty estimates are crucial for comparing experimental numbers. Are the measurements 0.86 s and 0.98 s the same or different? The answer depends on how exact these two numbers are. If the uncertainty too large, it is impossible to say whether the difference between the two numbers is real or just due to sloppy measurements. That's why estimating uncertainty is so important!

<table>
<thead>
<tr>
<th>Measurements don't agree</th>
<th>0.86 s ± 0.02 s and 0.98 s ± 0.02 s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurements agree</td>
<td>0.86 s ± 0.08 s and 0.98 s ± 0.08 s</td>
</tr>
</tbody>
</table>

If the ranges of two measured values don't overlap, the measurements are _discrepant_ (the two numbers don't agree). If the ranges overlap, the measurements are said to be _consistent_.

Understanding the difference between accuracy and precision

Accuracy and Precision Defined

**Accuracy** is the closeness of agreement between a measured measurement and a true or accepted value of the quantity being measured. **Precision** is a measure of how well a result can be determined. It is the degree of consistency and agreement among independent measurements of the same quantity. *Notice that a measurement can be precise without being accurate. (see analogy below)*

A Useful Analogy

Golf is a game of consistency and aim. Golfers practice to be able to put the ball in the cup (or as close as they can get) time after time. Four different golfers hit three shots each. The outcomes are shown below:

<table>
<thead>
<tr>
<th>Golfer A</th>
<th>Golfer B</th>
<th>Golfer C</th>
<th>Golfer D</th>
</tr>
</thead>
<tbody>
<tr>
<td>consistent, but off target</td>
<td>less consistent than Golfer A, but a little more on target.</td>
<td>the ideal golfer: consistent and on target</td>
<td>all over the place!</td>
</tr>
</tbody>
</table>

**Precision:** Good  
**Accuracy:** Poor  

**Precision:** OK  
**Accuracy:** Better  

**Precision:** Good  
**Accuracy:** Best  

**Precision:** Poor  
**Accuracy:** ???

*This analogy has one major flaw: *In science (unlike golf), the experimenter usually does not know where the target is!* A scientist feels a result is accurate when the results of the experiment agree with the results from several other experiments conducted by other scientists.*
An Example: The Golden Ring

Suppose you want to measure the mass of a gold ring that you want to sell to a friend so you can give him a fair price. Consider the three methods for determining the ring's mass:

- **Method 1:** Holding the ring in your hand, you estimate that the ring has a mass of somewhere between 10 g and 20 g.
- **Method 2:** You find an electronic balance which gives a mass reading of 17.43 grams. To make sure the balance is reliable, you use the same electronic balance and obtain several more readings: 17.46 g, 17.42 g, 17.44 g, so that the average mass appears to be in the range of 17.44 ± 0.02 g.
- **Method 3:** Since you want to be honest, you decide to use another balance. Again you do multiple trials. This balance gives readings of 17.22 g, 17.21 g, 17.21 g and 17.21 g.

Questions:

1. Which of these methods is most precise?

   Method 3 is the most precise. While both balances can be read to the nearest 0.01 g, the second balance appears to give a more consistent value.

2. Which one of these methods is most accurate?

   The only way to assess the accuracy of the measurement is to compare with a known standard. In this situation, it may be possible to calibrate the balances with a standard mass that is accurate within a narrow tolerance and is traceable to a primary mass standard at the National Institute of Standards and Technology (NIST). Calibrating the balances should eliminate the discrepancy between the readings and provide a more accurate mass measurement.

   (Note: The imprecision of Method 1 makes accuracy of that measurement irrelevant. While it's likely that the actual mass of the ring lies between 10 g and 20 g, Method 1 is not precise enough to get a usable estimate of the ring's mass.)
Random Error and Systematic Error

Definitions

All experimental uncertainty is due to either random errors or systematic errors. Random errors are statistical fluctuations (in either direction) in the measured data due to the precision limitations of the measurement device. Random errors usually result from the experimenter's inability to take the same measurement in exactly the same way to get exact the same number. Systematic errors, by contrast, are reproducible inaccuracies that are consistently in the same direction. Systematic errors are often due to a problem which persists throughout the entire experiment.

Note that systematic and random errors refer to problems associated with making measurements. Mistakes made in the calculations or in reading the instrument are not considered in error analysis. It is assumed that the experimenters are careful and competent!

How to minimize experimental error: some examples

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Example</th>
<th>How to minimize it</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random errors</td>
<td>You measure the mass of a ring three times using the same balance and get slightly different values: 17.46 g, 17.42 g, 17.44 g</td>
<td>Take more data. Random errors can be evaluated through statistical analysis and can be reduced by averaging over a large number of observations.</td>
</tr>
<tr>
<td>Systematic errors</td>
<td>The cloth tape measure that you use to measure the length of an object had been stretched out from years of use. (As a result, all of your length measurements were too small.)</td>
<td>Systematic errors are difficult to detect and cannot be analyzed statistically, because all of the data is off in the same direction (either too high or too low). Spotting and correcting for systematic error takes a lot of care.</td>
</tr>
</tbody>
</table>
|                 | The electronic scale you use reads 0.05 g too high for all your mass measurements (because it is improperly tared throughout your experiment). | • How would you compensate for the incorrect results of using the stretched out tape measure?  
• How would you correct the measurements from improperly tared scale? |
Common Sources of Error in Physics Lab Experiments

There is no such thing as "human error"! This vague phrase does not describe the source of error clearly. Careful description of sources of error allows future experimenters to improve on your techniques. This long list of common sources of error is meant to help you identify some of the common sources of error you might encounter while doing experiments. If you find yourself stuck for words when describing sources of error, this list may help. The list goes from the common to the obscure.

Incomplete definition (may be systematic or random) - One reason that it is impossible to make exact measurements is that the measurement is not always clearly defined. For example, if two different people measure the length of the same rope, they would probably get different results because each person may stretch the rope with a different tension. The best way to minimize definition errors is to carefully consider and specify the conditions that could affect the measurement.

Failure to account for a factor (usually systematic) - The most challenging part of designing an experiment is trying to control or account for all possible factors except the one independent variable that is being analyzed. For instance, you may inadvertently ignore air resistance when measuring free-fall acceleration, or you may fail to account for the effect of the Earth's magnetic field when measuring the field of a small magnet. The best way to account for these sources of error is to brainstorm with your peers about all the factors that could possibly affect your result. This brainstorm should be done before beginning the experiment so that arrangements can be made to account for the confounding factors before taking data. Sometimes a correction can be applied to a result after taking data, but this is inefficient and not always possible.

Environmental factors (systematic or random) - Be aware of errors introduced by your immediate working environment. You may need to take account for or protect your experiment from vibrations, drafts, changes in temperature, electronic noise or other effects from nearby apparatus.

Instrument resolution (random) - All instruments have finite precision that limits the ability to resolve small measurement differences. For instance, a meter stick cannot distinguish distances to a precision much better than about half of its smallest scale division (0.5 mm in this case). One of the best ways to obtain more precise measurements is to use a null difference method instead of measuring a quantity directly. Null or balance methods involve using instrumentation to measure the difference between two similar quantities, one of which is known very accurately and is adjustable. The adjustable reference quantity is varied until the difference is reduced to zero. The two quantities are then balanced and the magnitude of the unknown quantity can be found by comparison with the reference sample.
With this method, problems of source instability are eliminated, and the measuring instrument can be very sensitive and does not even need a scale.

**Failure to calibrate or check zero of instrument** (systematic) - Whenever possible, the calibration of an instrument should be checked before taking data. If a calibration standard is not available, the accuracy of the instrument should be checked by comparing with another instrument that is at least as precise, or by consulting the technical data provided by the manufacturer. When making a measurement with a micrometer, electronic balance, or an electrical meter, always check the zero reading first. Re-zero the instrument if possible, or measure the displacement of the zero reading from the true zero and correct any measurements accordingly. It is a good idea to check the zero reading throughout the experiment.

**Physical variations** (random) - It is always wise to obtain multiple measurements over the entire range being investigated. Doing so often reveals variations that might otherwise go undetected. If desired, these variations may be cause for closer examination, or they may be combined to find an average value.

**Parallax** (systematic or random) - This error can occur whenever there is some distance between the measuring scale and the indicator used to obtain a measurement. If the observer's eye is not squarely aligned with the pointer and scale, the reading may be too high or low (some analog meters have mirrors to help with this alignment).

**Instrument drift** (systematic) - Most electronic instruments have readings that drift over time. The amount of drift is generally not a concern, but occasionally this source of error can be significant and should be considered.

**Lag time and hysteresis** (systematic) - Some measuring devices require time to reach equilibrium, and taking a measurement before the instrument is stable will result in a measurement that is generally too low. The most common example is taking temperature readings with a thermometer that has not reached thermal equilibrium with its environment. A similar effect is hysteresis where the instrument readings lag behind and appear to have a "memory" effect as data are taken sequentially moving up or down through a range of values. Hysteresis is most commonly associated with materials that become magnetized when a changing magnetic field is applied.
Estimating uncertainty from a single measurement

In many circumstances, a single measurement of a quantity is often sufficient for the purposes of the measurement being taken. But if you only take one measurement, how can you estimate the uncertainty in that measurement? **Estimating the uncertainty in a single measurement requires judgement on the part of the experimenter.** The uncertainty of a single measurement is limited by the precision and accuracy of the measuring instrument, along with any other factors that might affect the ability of the experimenter to make the measurement and it is up to the experimenter to estimate the uncertainty (see the examples below).

---

**Example**

Try measuring the diameter of a tennis ball using the meter stick. What is the uncertainty in this measurement?

Even though the meterstick can be read to the nearest 0.1 cm, you probably cannot determine the diameter to the nearest 0.1 cm.

- What factors limit your ability to determine the diameter of the ball?
- What is a more realistic estimate of the uncertainty in your measurement of the diameter of the ball?

**Answers:** It's hard to line up the edge of the ball with the marks on the ruler and the picture is blurry. Even though there are markings on the ruler for every 0.1 cm, only the markings at each 0.5 cm show up clearly. I figure I can reliably measure where the edge of the tennis ball is to within about half of one of these markings, or about 0.2 cm. The left edge is at about 50.2 cm and the right edge is at about 56.5 cm, so the diameter of the ball is about 6.3 cm ± 0.2 cm.
Another example

Try determining the thickness of a CD case from this picture.

- How can you get the most precise measurement of the thickness of a single CD case from this picture? (Even though the ruler is blurry, you can determine the thickness of a single case to within less than 0.1 cm.)
- Use the method you just described to determine the thickness of a single case (and the uncertainty in that measurement)
- What implicit assumption(s) are you making about the CD cases?

Answers: The best way to do the measurement is to measure the thickness of the stack and divide by the number of cases in the stack. That way, the uncertainty in the measurement is spread out over all 36 CD cases. It's hard to read the ruler in the picture any closer than within about 0.2 cm (see previous example). The stack goes starts at about the 16.5 cm mark and ends at about the 54.5 cm mark, so the stack is about 38.0 cm ± 0.2 cm long. Divide the length of the stack by the number of CD cases in the stack (36) to get the thickness of a single case: 1.056 cm ± 0.006 cm. By "spreading out" the uncertainty over the entire stack of cases, you can get a measurement that is more precise than what can be determined by measuring just one of the cases with the same ruler. We are assuming that all the cases are the same thickness and that there is no space between any of the cases.
Estimating uncertainty from multiple measurements

Increasing precision with multiple measurements

One way to increase your confidence in experimental data is to repeat the same measurement many times. For example, one way to estimate the amount of time it takes something to happen is to simply time it once with a stopwatch. You can decrease the uncertainty in this estimate by making this same measurement multiple times and taking the average. The more measurements you take (provided there is no problem with the clock!), the better your estimate will be.

Taking multiple measurements also allows you to better estimate the uncertainty in your measurements by checking how reproducible the measurements are. How precise your estimate of the time is depends on the spread of the measurements (often measured using a statistic called standard deviation) and the number (N) of repeated measurements you take.

Consider the following example: Maria timed how long it takes for a steel ball to fall from top of a table to the floor using the same stopwatch. She got the following data:

0.32 s, 0.54 s, 0.44 s, 0.29 s, 0.48 s

By taking five measurements, Maria has significantly decreased the uncertainty in the time measurement. Maria also has a crude estimate of the uncertainty in her data; it is very likely that the "true" time it takes the ball to fall is somewhere between 0.29 s and 0.54 s. Statistics is required to get a more sophisticated estimate of the uncertainty.
Some statistical concepts

When dealing with repeated measurements, there are three important statistical quantities: average (or mean), standard deviation, and standard error. These are summarized in the table below:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>What it is</th>
<th>Statistical interpretation</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>an estimate of the &quot;true&quot; value of the measurement</td>
<td>the central value</td>
<td>$x_{\text{ave}}$</td>
</tr>
<tr>
<td>standard deviation</td>
<td>a measure of the &quot;spread&quot; in the data</td>
<td>You can be reasonably sure (about 70% sure) that if you repeat the same measurement one more time, that next measurement will be less than one standard deviation away from the average.</td>
<td>$s$</td>
</tr>
<tr>
<td>standard error</td>
<td>an estimate in the uncertainty in the average of the measurements</td>
<td>You can be reasonably sure (about 70% sure) that if you do the entire experiment again with the same number of repetitions, the average value from the new experiment will be less than one standard error away from the average value from this experiment.</td>
<td>$SE$</td>
</tr>
</tbody>
</table>

Maria's data revisited

The statistics for Maria's stopwatch data are given below:

- $x_{\text{ave}} = 0.41$ s
- $s = 0.11$ s
- $SE = 0.05$ s

It's pretty clear what the average means, but what do the other statistics say about Maria's data?

- **Standard deviation**: If Maria timed the object's fall once more, there is a good chance (about 70%) that the stopwatch reading she will get will be within one standard deviation of the average. In other words, the next time she measures the
time of the fall there is about a 70% chance that the stopwatch reading she gets will be between (0.41 s - 0.11 s) and (0.41 s + 0.11 s).

- **Standard error**: If Maria did the entire experiment (all five measurements) over again, there is a good chance (about 70%) that the average of the those five new measurements will be within one standard error of the average. In other words, the next time Maria repeats all five measurements, the average she will get will be between (0.41 s - 0.05 s) and (0.41 s + 0.05 s).
Calculating the statistics using Excel

Spreadsheet programs (like Microsoft Excel) can calculate statistics easily. Once you have the data in Excel, you can use the built-in statistics package to calculate the average and the standard deviation.

To calculate the average of cells A4 through A8:

- Select the cell you want the average to appear in (D1 in this example)
- Type "=average(a4:a8)"
- Press the Enter key

To calculate the standard deviation of the five numbers, use Excel's built-in STDEV function.
Excel doesn't have a **standard error** function, so you need to use the formula for standard error:

$$SE = \frac{s}{\sqrt{N}}$$

where $N$ is the number of observations.
Estimating uncertainty in a calculated quantity

Often we calculate quantities from measurements we make. Since the measurements are not exact, quantities calculated from the measurements are not exact either. This page shows a method for figuring out the uncertainty in a calculated quantity from the uncertainties in the measurement(s).

Upper Bound / Lower Bound Method

The Upper Bound / Lower Bound Method is a simple way to get a rough estimate of the uncertainty in a calculated quantity. The basic procedure is

1. Plug in the measurements into the equation so that the value for the calculated quantity is as large as it can possibly be. (This is the upper bound).
2. Plug in the measurements into the equation so that the value for the calculated quantity is as small as it can possibly be. (This is the lower bound).
3. To calculate the value, take the average of the upper and lower bound.
4. To find the uncertainty, simply take the difference between the upper and lower bounds and divide by 2.
5. Round off the results appropriately.

This method is crude, but works easily with any calculation. More sophisticated methods for estimating the uncertainty in a calculation involve calculus and lead to somewhat lower uncertainty estimates.

Example: Prove it, Officer!

Cars on Dixie Trail often exceed the 35 mph speed limit in front of your house by 10 mph or more, especially during rush hour. As concerned parents of two kids, armed only with a stopwatch and a tape measure, you and your husband set out to catch some speeders. You carefully measure the distance between two trees (235 ft ± 3 ft). You time the cars as your husband writes down the license plate numbers. One of the cars you time takes 3.85 s to get from one tree to the next. Can you prove that this driver was speeding? (Assume that the uncertainty in the time measurement is 0.1 s.)
1. The highest possible value (the Upper Bound) for the speed of the car from your measurements is:

\[ s_{\text{upper}} = \frac{d}{t} = \frac{(238 \, \text{ft})}{(3.75 \, \text{s})} = 63.47 \, \text{ft/s} \times \frac{3600 \, \text{s}}{1 \, \text{hr}} \times \frac{1 \, \text{mi}}{5280 \, \text{ft}} = 43.27 \, \text{mph} \]

**Note:** To make the fraction as large as possible, the longest distance (285+3) ft must be divided by the shortest time (0.85-0.1) s.

2. The lowest possible value (the Lower Bound) for the speed of the car from your measurements is:

\[ s_{\text{lower}} = \frac{d}{t} = \frac{(232 \, \text{ft})}{(3.95 \, \text{s})} = 58.73 \, \text{ft/s} \times \frac{3600 \, \text{s}}{1 \, \text{hr}} \times \frac{1 \, \text{mi}}{5280 \, \text{ft}} = 40.05 \, \text{mph} \]

3. ...so the speed of the car is about \( (43.27 \, \text{mph} + 40.05 \, \text{mph})/2 = 41.66 \, \text{mph} \)
4. ...and the uncertainty in the speed is about \( (43.27 \, \text{mph} - 40.05 \, \text{mph})/2 = 1.61 \, \text{mph} \)
5. Now round off. Since your speed value has an uncertainty of more than 1 mph, it doesn't make sense to quote digits beyond the ones place. The car's speed is \( (42 \pm 2) \, \text{mph} \)

Despite your seemingly primitive measurements, you can PROVE that this car was speeding. Even your measurements had been off by as much as 3 ft and 0.1 s, the calculated speed of the car was no less than 40 mph (and no more than 44 mph).
# Common Student Mistakes

In the table below, you will find some examples of mistakes students often make in their lab reports. Read through the examples and see if you can spot the problem. Then read to see how to fix the problem.

<table>
<thead>
<tr>
<th>Example(s)</th>
<th>What's wrong</th>
<th>How to fix/avoid the mistake</th>
</tr>
</thead>
<tbody>
<tr>
<td>From the analysis, we concluded that increasing the amount of surface area</td>
<td>Without some estimate of the precision of both numbers, the reader cannot tell whether the difference between two numbers is due to something the experimenters changed on purpose or the inexactness of the measurements!</td>
<td>Always include estimates for the uncertainties in numbers you compare. Imagine how this paragraph might have read if the uncertainty in the force measurements was ±0.5 N. (In this case, the group would have reported the WRONG CONCLUSION!)</td>
</tr>
<tr>
<td>1.838 N for just the sled and 1.635 N for the sled on its side. [paraphrased from PY205 student lab report, Fall 2000]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;[...] and human error could have also played a part.&quot;</td>
<td>The phrase (&quot;human error&quot;) is vague. Sources of error should be described in enough detail that the report reader can tell how to improve the procedure.</td>
<td>Always be specific about sources of error. (For example: We had trouble starting and stopping the stopwatch at exactly the right time). If you are not sure how to describe the source of error, see the Sources of Error section of this guide for some examples. Don't use the phrase &quot;human error&quot;!</td>
</tr>
<tr>
<td>&quot;Our limitations in this experiment were [...] our group may have rushed to get the lab in on time&quot;</td>
<td>Mistakes made in the analysis are not considered sources of experimental error. Only problems associated with making the measurements are considered sources of experimental error.</td>
<td>Do not list mistakes (like mistyping the calculator, forgetting to take square root, etc.) as sources of error. Don't use the phrase &quot;human error&quot;!</td>
</tr>
<tr>
<td>&quot;The sources of error were [...] as well as human error in reading the numerical values on the graph.&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[quoted from two different PY205 student lab reports, Fall 2000]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.219 s ± 0.5 s</td>
<td>The measurement cannot have more decimal places than the error estimate. How can you say anything about the thousandths place (or the hundredths place, for that matter) if you are unsure of the tenths place?</td>
<td>Always round the measurement off to the same decimal place as the error estimate: 3.2 s ± 0.5 s</td>
</tr>
<tr>
<td>4258 nm ± 523 nm</td>
<td>Error estimates are rough estimates, so they should be rounded to one significant digit.*</td>
<td>Round the error estimate to one significant digit: 4200 nm ± 500 nm (The rounding on the measurement is adjusted accordingly: see example above)</td>
</tr>
</tbody>
</table>

* Some scientists occasionally round the error estimate to two significant digits if the error estimate begins with a one: 7.99823 MHz ± 0.00014 MHz.