Introductory Physics Students' Treatment of Measurement Uncertainty

by

Duane Lee Deardorff

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

2001

Approved by:

Dr. Robert Beichner (Chair of Advisory Committee)	Dr. John Risley
Dr. Jacquelin Dietz	Dr. John Park

It is better to be roughly right
than precisely wrong.

- Alan Greenspan, 1997

Chairman of the U.S. Federal Reserve Board

Biography

Duane Lee Deardorff was born in Quito, Ecuador on April 16, 1969, during his parents' missionary work that began there in 1967. When he was two months old, he moved with his parents, Darryl and Juanita, to Huber Heights, Ohio, and lived there until he went to college. While growing up, he became fascinated with flying objects and spent many hours designing, creating, and flying model rockets, airplanes, kites, and boomerangs. His interest in flying objects expanded as he learned to juggle at age twelve. Since then, he has performed hundreds of juggling shows for children of all ages - even once in a Hollywood movie!

Duane benefited from a Christian liberal-arts education as a student at Manchester College in Indiana. It was here that he drove his physics lab partner crazy trying to make precise measurements while learning to quantify the uncertainty in experimental results. After graduating from Manchester in 1991 with a B.A. in physics, he moved to Virginia and married Darla Kay Bowman, whom he had met two years earlier at a church conference in Florida. They lived in the Shenandoah Valley while Darla Kay completed her senior year at Bridgewater College. While in Virginia, Duane worked as an Environmental Scientist for Triad Engineering and also taught several physics labs at James Madison University. In 1993, Duane began his graduate work in physics at North Carolina State University. Although he had originally chosen NCSU to pursue research in condensed matter physics to improve photovoltaic technology, he realized that he had a deeper passion for teaching and educational reform efforts. Fortunately, the NCSU department of physics has one of the top physics education research groups, so Duane joined this new field of study and eagerly participated in the innovative physics education initiatives under the guidance of Dr. Robert Beichner.

In January 2000, Duane took a physics faculty position as Director of Undergraduate Laboratories at the University of North Carolina at Chapel Hill. Duane and Darla Kay moved from Raleigh to Durham, NC, only days before their first child, Kaylee, was born on April 24. Now Kaylee is teaching her parents to see the world from a whole new perspective.

Acknowledgements

I would like to acknowledge the cooperation and support of the universities and departments that helped make this research possible: NC State University department of physics, the department of physics and astronomy in the University of North Carolina at Chapel Hill, the University of Hokkaido in Sapporo, Japan, and Seoul National University in South Korea. The research in Japan and Korea was made possible in part from funding through a Dissertation Enhancement Grant from the National Science Foundation (NSF grant DUE-9752313). This international collaboration was facilitated by the gracious assistance of Dr. Sugiyama Shigeo who volunteered to coordinate my visit and delivery of my research surveys. His graduate student, Kaori Takaguchi, was also valuable as the translator during interviews with Japanese students. I also benefited greatly from my visit with Korean graduate student Jongah Soh, as she discussed with me her research that closely parallels mine.

Special thanks go to my advisor, Robert Beichner, for his kind support, encouragement, and guidance throughout my graduate school career. I have benefited from many valuable discussions with other members of the NCSU Physics Education Research

Group: Jeff Saul, Scott Bonham, David Abbott, Rhett Allain, and Melissa Dancy, as well as John Risley, Peg Gjertsen, and Larry Martin.

I express my greatest appreciation to my wife, Darla Kay Deardorff, for all the patience and sacrifice she gave while I spent hundreds of hours working on my research. I hope to return the favor soon as she prepares to write her dissertation. I am also grateful to my daughter, Kaylee, who helped me keep my perspective on what is most important in life, and also gave me the motivation to complete my degree.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Research Objective

The objective of this qualitative study is to examine and document how introductory physics students treat the uncertainty of measurements. In meeting this objective, the conceptions and practices of physics instructors (primarily graduate teaching assistants) are also examined in order to define a reference standard to which the student practices may be compared.

1.2 Research Questions

This study is guided by the following questions:

1. What are the common conceptions, practices, or "Facets" (Minstrell 1992) demonstrated by introductory physics students regarding measurement uncertainty and error analysis?
2. How do students treat the uncertainty in measurements differently than experts (graduate students, professors, and authors of reference materials)?
3. Why do students believe what they do about measurement uncertainty? (Answering this question helps facilitate the development of more effective curricular materials for teaching measurement uncertainty.)

1.3 Motivation for this Research

Despite the fact that extensive research efforts have been made in recent years to better understand how students learn physics, very few studies have addressed students' understanding of the inherent uncertainty associated with physical measurements (Sere, Journeaux et al. 1993; Lubben and Millar 1996; Allie, Buffler et al. 1998; Soh, Fairbrother et al. 1998). Of these studies, none have made a comprehensive effort to examine the scope of this concept, even though it is a critical component of all scientific investigations. As nearly all physics lab instructors can attest, introductory students (and even advanced students) often have difficulty understanding and analyzing the uncertainties in their measurements (see Chapter 2). Therefore, it seems prudent to investigate these difficulties and try to understand the cause for confusion and misunderstanding so that instruction on this subject can be improved. This subject is also worthy of investigation because it has many diverse applications in a variety of disciplines as explained in the following sections.

1.4 The Nature of Uncertainty in Measurement

Measurement uncertainty is an intrinsic part of all scientific investigations. Science is based on the systematic pursuit of knowledge involving the collection of data through observation and experiment, and the formulation and testing of hypotheses (Merriam-Webster 2000). The laws of nature as we know them have been developed and tested from years of scientific investigation. The process of scientific inquiry naturally leads to the important questions about how well an empirical result is known, whether or not the result agrees with a hypothesis or theoretical prediction, and whether the result can be verified by other researchers. In order to answer these basic questions, the uncertainty of the measured

result must be estimated and quantified to indicate the degree of confidence associated with the measurement. Only after the uncertainty of an experimental result is established can a reasonable conclusion be made about how the result compares with a theoretical prediction or some other experimental value. Therefore, the process of determining the uncertainty of measurements (commonly called *error analysis*) is fundamental to all scientific investigations.

Physics is the study of matter and energy interactions and is the most fundamental of the natural sciences. Nearly all of the physics principles taught to students are based on experimentation, and every experiment requires measurements that are inherently uncertain. Introductory physics laboratory courses provide a natural opportunity for students to learn the fundamental practices of experimentation and data analysis. As will be shown in this dissertation, these practices are not easy for students to master, but the effort to do so is worthwhile since the concepts have important applications in a variety of fields beyond physics.

The following quote summarizes the importance of improving instruction in the area of error analysis:

> It has been a considerable handicap to many experimenters that their formal scientific training has left them unequipped to deal with the common situation in which experimental error cannot be safely ignored. Not only is awareness of the possible effects of experimental error essential in the analysis of data, but also its influence is a paramount consideration in *planning* the generation of data, that is, in the *design* of experiments. Therefore, to have a sound base on which to build practical techniques for the design and analysis of experiments, some elementary understanding of experimental error and of associated probability theory is essential (Box, 1978, p.24).

1.5 Student Difficulties with Measurement Uncertainty

The primary reason for investigating student treatment of measurement uncertainty is that there is widespread anecdotal evidence from physics teachers that students have difficulty analyzing measurement errors. A goal of this research is to determine how widespread these misunderstandings really are, and whether the situation is as bad as teachers believe. Below are some common student behaviors that have been observed by the author and other physics instructors (a more comprehensive list can be found in Chapter 2):

- Students often fail to consider the uncertainty in measured values when evaluating whether two results are in agreement.
- Students apply rules of significant figures without a firm conceptual understanding of why they are used.
- Students frequently report results with more (in)significant figures than can be justified.
- Student comments indicate limited thought about the nature of uncertainty:
  "We used a computer to analyze our data so there was no error in our result."
  "The primary source of error in our experiment was *human error*."

Concern about these behaviors contributed to the motivation for this research. A more extensive investigation of the student learning objectives targeted by this study is presented in Chapter 2.

1.6 Applications in Physics

Precision measurements are inherently linked to physics, especially in experiments designed to push the limits of what we know about the physical world. This is the reason that the National Institute of Standards and Technology (NIST) has a physics division that is responsible in part for continuing to measure and report the fundamental physical constants to the greatest precision possible. NIST is a federal agency of the U.S. Department of

Commerce and is responsible for communicating with industry to develop and apply technology, measurements, and standards. To facilitate this communication, NIST has published guidelines for evaluating and reporting the uncertainty of measurements (Taylor and Kuyatt 1994).

1.7 Applications Beyond Physics

One reason for investigating student understanding of measurement uncertainty is that it is a truly fundamental concept that has applications in many diverse scientific fields including metrology (the study of measurements), statistics, physics, engineering, chemistry, economics, and even the social sciences. The use of scientific data is certainly not limited to researchers, laboratory technicians, and engineers. The general public is also responsible for interpreting scientific reports and making decisions based on the results of experimental studies. Unfortunately, many members of society are numerically illiterate and do not have the skills necessary to make sound decisions despite all the quantitative data that are available to them. In his book *Innumeracy*, John Paulos gives numerous examples of situations where people often do not use or understand numerical data and the consequences this misunderstanding can have on their lives (Paulos 1988). Students who learn data analysis skills (in a physics lab or by some other means) should be better prepared to make sense of numerical data they encounter in both their careers and personal lives. This viewpoint is supported by an excerpt taken from a 1997 paper presented by Maryanne Fox at a symposium in Washington, D.C., held by the Center for Science, Mathematics, and Engineering Education to reflect on educational reform during the past 40 years since Sputnik:

> As scientists, mathematicians, and engineers, many of us are completely astonished by our students' inability to understand scale. One of my colleagues asked his freshman students this fall to estimate the diameter of the earth. From a class of several hundred, he got two responses: 100 miles and 1.41 million miles. The first student had just arrived in Austin from Waco, a distance of about 100 miles, and perhaps the distance from home to college did represent the ends of the earth to him. But the second one? How can one be so wrong with such precision? How bewildering living every day within nature must be to such students?(Fox 1997)

Every measured value has some degree of uncertainty, and while most circumstances do not warrant an extensive error analysis, there are numerous situations beyond the field of physics where a reasonably accurate determination of the uncertainty is important for making critical decisions. Several important examples are provided in the following sections.

1.7.1 Legal Decisions

Many important legal decisions depend on the accuracy of scientific data that is inherently subject to uncertainty. Consequently, the Federal Rules of Evidence for court testimony by experts and technical data were revised in 1993 to include the following points: (Bernstein 1993).

1) The court should determine whether the theory or technique in question can be (or has been) tested.
2) Peer review is an important consideration.
3) The known or potential rate of error of the technique should be determined, as should the existence and maintenance of standards controlling the technique's operation.

The Federal Rules of Evidence place the responsibility on attorneys to validate the accuracy of any scientific evidence presented in court cases. However, jurors should also have a reasonable understanding of the nature of errors associated with empirical data so that they

can make well-educated decisions, especially since their judgments will affect the future of other human beings.

1.7.2 Environmental Risk Assessment

One specific application of measurement uncertainty is in the assessment of environmental risks for human health and safety. The United States Environmental Protection Agency (EPA) often requires scientific testing to determine if environmental contamination levels are below a safe limit. These tests must be precise enough to examine concentration levels at or below critical exposure levels. A test that is not sufficiently precise cannot be used to make a reliable judgment, since the measurement result of "ND" (none detected) can give a false sense of security.

An important environmental example that has significant global and economic consequences is the issue of global warming. Scientists have been examining the possibility that the average temperature of the Earth is rising, which if unchecked, could result in devastating flooding of large areas from excessive melting of the polar ice caps. As with many scientific investigations, the data that can be obtained and analyzed are limited, and while there appears to be an overall warming trend over the past few decades, the variability and uncertainty in the data must also be considered in any conclusions that are made (Jones and Wigley 1990).

1.7.3 Economic Forecasting

The level of uncertainty in measurements is especially critical when trying to predict future activity by extrapolating from current and past data. Economic forecasting is used by financial and business planners in an attempt to predict future financial figures based on

historical patterns. The amount of uncertainty in these predictions can significantly affect the decisions of investors and financial officers. Federal Reserve Board Chairman, Alan Greenspan, summed up the nature of uncertainty in economic forecasting by quoting the British economist John Maynard Keynes, who said, "It is better to be roughly right, than precisely wrong" (NPR 1997). This quote (by a man whose words are heavily weighted) suggests that uncertainty in an estimate is acceptable, as long as all known systematic errors have been eliminated so that the estimate is (hopefully) centered on the target value (see Figure 4-1). This statement also summarizes the expert perspective on measurement uncertainty – reasonably accurate results are more beneficial than precise results that have no validity.

1.7.4 Weather Forecasting

Weather forecasting is one of the most common examples where the uncertainty of a prediction is reported (e.g., “The chance of rain tomorrow is 80%”). Even with high-tech meteorological equipment there are no guarantees in predicting future weather conditions. Although weather forecasts are one of the few examples where an explicit probability is often reported for a prediction, they are familiar to almost everyone. Despite frequent exposure to weather forecasts where the probability is almost always rounded to the nearest 5 or 10%, some students insist on reporting unreasonably precise experimental values equivalent to stating that the chance of rain is 80.537%. Evidently, simple exposure to correct reporting methods is not sufficient for students to recognize the purpose of significant figures.

1.7.5 Public Opinion Polls

Results from public opinion polls are another one of the few instances where the margin of error is regularly reported in data that are presented to the general public. Prior to taking a physics or chemistry class, the fine print beneath these poll results (e.g., the margin of error is ± 3%) may be the only exposure students have to notation that explicitly shows the relative uncertainty of a measurement (based on responses from student interviews). Unfortunately, this ± notation for margin of error represents a different confidence interval than is typically used in physics (see

Table 2-1).

1.7.6 Quality Assurance and Control

Uncertainty estimates play a critical role in quality assurance and control processes. Statistical analyses form the basis of many of the decisions made in these areas. In fact, the ISO 9000 industry standards for quality assurance require that test measurements include an estimate of their uncertainty as specified in the International Standards Organization (ISO) *Guide to the Expression of Uncertainty in Measurement* (ISO 1993). ISO 9000 standards state:

> The supplier shall determine the measurements to be made and the accuracy required, and select the appropriate inspection, measuring, and test equipment that is capable of the necessary accuracy and precision. Inspection, measuring, and test equipment shall be used in a manner which ensures that the measurement uncertainty is known and is consistent with the required measurement capability.

Students who intend to pursue careers in industrial engineering or manufacturing quality control could be better prepared by learning the fundamentals of measurement uncertainty in an introductory physics course.

1.8 Physics Education Research

Physics education research is a relatively new academic discipline, but one that is growing quickly. Over the last twenty-five years, an increasing number of physicists, science education researchers, and cognition specialists have been carefully examining how students learn physics. These researchers have succeeded in uncovering many student misconceptions and the reasoning that underlies these conceptual difficulties. Their research findings have been used to develop new curricula that intentionally address these difficulties, and which have been shown to dramatically improve students' fundamental understanding of physics concepts. The research for this dissertation is similar to other physics education studies that have investigated students' understanding in specific content areas. Table 1-1 lists the content areas that have been examined in the last two decades, along with an estimate of the number of studies for each topic. This list is a tally of the studies included in a 1998 Resource Letter on Physics Education Research (McDermott and Redish 1998). While this list is not meant to be exhaustive, it at least provides insight into the relative emphasis that researchers have placed on various subjects.

Table 1-1. Recent empirical studies in physics education research

Content Area	Studies
Mechanics	
Kinematics	8
Dynamics	18
Relativity and reference frames	5

Electricity and magnetism	
DC circuits	10
Electrostatics and magnetostatics	2
Electric and magnetic fields	6
Light and optics	
Nature of light, color, and vision	5
Geometrical optics	4
Physical optics	1
Properties of matter, thermal physics	
Heat, temperature, and thermodynamics	10
Pressure, density, and the structure of matter	4
Waves and sound	4
Modern physics	3
Problem-solving performance	4
Laboratory instruction and demonstrations	5
Ability to apply mathematics in physics	4
Attitudes and beliefs of students	11
Student reasoning	4

Despite their fundamental importance, the topics of measurement and precision have largely been ignored by physics education researchers, even though these areas are generally addressed in the very first chapter of most physics textbooks. One possible reason for this deficiency is that measurement practices are generally covered in the laboratory section of introductory physics courses, and most physics education research has focused on the mainstream curriculum, as evidenced by the relatively few studies related to laboratory instruction (only 4 of the 108 studies listed in Table 1-1). In fact, only one of the studies (Sere, 1993) addresses students' conceptual understanding of measurements. The purpose of this dissertation research is to begin to fill this gap in understanding how physics students think about the accuracy of measurements they make.

Several educators have expressed their concern that procedural knowledge taught in labs has been de-emphasized relative to declarative knowledge taught in lectures and

tutorials (Swartz 1995; Osborne 1996; Allie, Buffler et al. 1998). This study addresses that concern by examining the ability of students to make accurate measurements, estimate the uncertainty in those measurements, evaluate the quality of their results, and design experiments based on the degree of precision required. This investigation will help establish a research base on procedural knowledge for experimentation. From this foundation, instructors can develop their curricula to better address the needs and learning difficulties of their students.

1.9 Summary

The principal objective of this research is to examine and document introductory physics students' conceptions and practices related to measurement uncertainty. Surprisingly little research has been conducted to examine students' understanding of these topics, despite the fact that measurements and standards are usually addressed in the first chapter of most physics books and many other physics topics have already been investigated. Expert knowledge (from reference materials and surveys) will serve as a standard to which the student performance will be compared. The ultimate goal of this research is then to provide the physics education community with useful information that can facilitate curriculum development and improved instruction on this fundamental topic.

2 Background

2.1 Definitions of Terms

Terminology and notation related to measurement uncertainty is not used consistently among experts. In order to clarify the meaning of terms used in this dissertation, and to show the range of meanings, a compilation of key terms with definitions is included here. The definitions are taken from a sample of reference sources that represent the scope of this study (the three most popular reference books in Table 2-2, plus the ISO Guide and an industrial metrology reference book). Definitions from Webster's dictionary are also included for several of the terms to show the contrast between common vernacular use and the specific meanings of these terms as they relate to scientific measurements.

Sources:

- Taylor, John. *An Introduction to Error Analysis: The study of uncertainties in physical measurements,* 2nd. ed. University Science Books: Sausalito, CA, 1997.
- Bevington, Phillip R. and D. Keith Robinson. *Data Reduction and Error Analysis for the Physical Sciences,* 2nd. ed. McGraw-Hill: New York, 1992.
- Baird, D.C. *Experimentation: An Introduction to Measurement Theory and Experiment Design,* 3rd. ed. Prentice Hall: Englewood Cliffs, NJ, 1995.
- ISO. *Guide to the Expression of Uncertainty in Measurement.* International Organization for Standardization (ISO) and the International Committee on Weights and Measures (CIPM): Switzerland, 1993.
- Fluke. *Calibration: Philosophy and Practice,* 2nd. ed. Fluke Corporation: Everett, WA, 1994.
- *Webster's Tenth New Collegiate Dictionary*, Merriam-Webster: Springfield, MA, 2000.

Notes: The definitions presented below are provided to explain the meanings of terms used in this dissertation, and are therefore organized according to their meaning rather than an alphabetized list. Many of these terms are defined in the *International Vocabulary of Basic and General Terms in Metrology* (abbreviated VIM), and their identification numbers are shown in brackets immediately after the term (ISO 1993). Since the meaning and usage of these terms are not consistent among other references, alternative (and sometimes conflicting) definitions are provided with the name and page number of the reference from the above list. Comments are included in *italics* to elaborate on several of the definitions. References are only cited when they explicitly define a term. Omission of a reference for a particular term generally indicates that the term was not used or clearly defined by that reference. Even more diverse usage of these terms exists in other references not cited here.

> **uncertainty (of measurement)** [VIM 3.9] – **1.** 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). **2.** An estimate of the error in a measurement, often stated as a range of values that contain the true value within a certain confidence level (usually ± 1 σ for 68% confidence interval) (Taylor, p. 14; Fluke, p. G-15). **3.** Based on either limitations of the measuring instruments or from statistical fluctuations in the quantity being measured (Baird, p. 2). **4.** Indicates the precision of a measurement (Bevington, p. 2). (*All but this last definition suggest that the uncertainty includes an estimate of the precision **and** accuracy of the measured value*.)
>
> **(absolute) uncertainty** – **1.** the amount (often stated in the form ± δx) that along with the measured value, indicates the range in which the desired or true value most likely lies (Baird, p. 14). **2.** The total uncertainty of a value (Fluke, p. G-3). **3.** The error (Taylor, p. 14). (*Taylor does not distinguish between the terms **error** and **uncertainty**, which is an unfortunate source of confusion for anyone who refers to this popular book.*)

relative (fractional) uncertainty – the absolute uncertainty divided by the measured value, often expressed as a percentage or in parts per million (ppm) (Taylor, p. 28; Baird, p. 14).

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

combined standard uncertainty, $u_c(y)$ – the standard deviation of the result of a measurement when the result is obtained from the values of a number of other quantities. It is obtained by combining the individual standard uncertainties u_i (and covariances as appropriate), using the law of propagation of uncertainties, commonly called the "root-sum-of-squares" or "RSS" method. The combined standard uncertainty is commonly used for reporting fundamental constants, metrological research, and international comparisons of realizations of SI units (ISO, p. 3).

Type A evaluation of standard uncertainty – method of evaluation of uncertainty by the statistical analysis of a series of observations (ISO, p. 3).

Type B evaluation of standard uncertainty – method of evaluation of uncertainty by means other than the statistical analysis of series of observations (ISO, p. 3).

precision – **1.** the degree of consistency and agreement among independent measurements of a quantity under the same conditions (Fluke, p. G-11). **2.** Indicated by the uncertainty (Bevington, p. 2), or **3.** the fractional (relative) uncertainty (Taylor, p. 28). **4.** The degree of refinement with which an operation is performed or a measurement stated (Webster). *Precision is a measure of how well the result has been determined (without reference to a theoretical or true value), and the reproducibility or reliability of the result. The fineness of scale of a measuring device generally affects the consistency of repeated measurements, and therefore, the precision. The ISO has banned the term* ***precision*** *for describing scientific measuring instruments because of its many confusing everyday connotations* (Giordano 1997).

accuracy (of measurement) [VIM 3.5] – **1.** closeness of agreement between a measured value and a true value (ISO, p. 33; Fluke, p. G-3; Bevington, p. 2; Taylor, p. 95). **2.** The term "precision" should not be used for "accuracy" (ISO, p. 33). **3.** A given accuracy implies an equivalent precision (Bevington, p. 3). **4.** Freedom from mistake or error, correctness; degree of conformity of a measure to a standard or a true value (Webster).

true value (of a quantity) [VIM 1.19] – **1.** 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). **2.** The correct value of the measurand (Fluke, p. G-15). **3.** The value that is approached by averaging an increasing number of measurements with no systematic errors (Taylor, p. 130).

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). (*This distinction is not clear in other references that refer to "the true value" of a quantity*.)

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).

error (of measurement) [VIM 3.10] – **1.** 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). **2.** Deviation from the "true" or nominal value (Bevington, p. 5; Fluke, p. G-7). **3.** The inevitable uncertainty inherent in measurements, not to be confused with a *mistake* or *blunder* (Taylor, 3). **4.** The amount of deviation from a standard or specification; **5.** mistake or blunder (Webster). (*Students often cite "human error" as a source of experimental error, and the dictionary definition of the term error only confuses this misused term. Here again, Taylor does not distinguish between the terms* ***error*** *and* ***uncertainty****, which clearly have different meanings according to the ISO.)*

random error [VIM 3.13] – **1.** 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). **2.** Statistical fluctuations (in either direction) in the measured data due to the precision limitations of the measurement device (Fluke, p. G-12; Taylor, p. 94).

systematic error [VIM 3.14] – **1.** 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). **2.** A reproducible discrepancy between the result and "true" value that is consistently in the same direction (Baird, p. 14; Fluke, p. G-14). **3.** A reproducible inaccuracy introduced by faulty equipment, calibration, or technique (Bevington, p. 3, 14). **4.** These errors are difficult to detect and cannot be analyzed statistically (Taylor, p. 11). **5.** Systematic error is sometimes called "bias" and can be reduced by applying a "correction" or "correction factor" to compensate for an effect recognized when calibrating against a standard. Unlike random errors, systematic errors cannot be reduced by increasing the number of observations (ISO, p. 5).

mistake or **blunder** – a procedural error that should be avoided by careful attention (Taylor, p. 3). These are illegitimate errors and can generally be corrected by carefully repeating the operations (Bevington, p. 2).

discrepancy – a significant difference between two measured values of the same quantity (Taylor, p. 17; Bevington, p. 5). (*Neither of these references clearly defines what is meant by a "significant difference," but the implication is that the difference*

between the measured values is clearly greater than the combined experimental uncertainty.)

relative error [VIM 3.12] – error of measurement divided by a true value of the measurand (ISO, p. 34). (*Relative error is often reported as a percentage. The relative or "percent error" could be 0% if the measured result happens to coincide with the expected value, but such a statement suggests that somehow a perfect measurement was made. Therefore, a statement of the uncertainty is also necessary to properly convey the quality of the measurement*.)

significant figures – all digits between and including the first non-zero digit from the left, through the last digit (Bevington, p. 4). (e.g., *0.05070 has 4 significant figures*.)

decimal places – the number of digits to the right of the decimal point. (*This term is not explicitly defined in any of the examined references*.)

sample standard deviation – the positive square root of the sample variance (see standard error)

standard error (standard deviation of the mean) – the sample standard deviation divided by the square root of the number of observations:

$$\text{SE} = s_{\bar{x}} = \frac{s}{\sqrt{n}} \quad \text{where } s^2 = \sum_i \frac{(x_i - \bar{x})^2}{n - 1} \text{ is the sample variance (ISO, p.38).}$$

Random errors are reduced by averaging over a large number of observations, because the standard error decreases as the sample size n increases (Taylor, p. 103).

(Note: The ISO Guide and most statistics books use the letter s to represent the sample standard deviation and σ (sigma) to represent the standard deviation of the population; however, σ is often used in casual error analysis discussions to indicate the sample standard deviation.)

margin of error – range of uncertainty. Public opinion polls generally use *margin of error* to indicate a 95% confidence interval, corresponding to an uncertainty range of $x \pm 2\sigma$ (Taylor, p. 14).

tolerance – the limits of the range of values (the uncertainty) that apply to a properly functioning measuring instrument (Fluke, p. 3-7).

coverage factor, k – numerical factor used as a multiplier of the combined standard uncertainty in order to obtain an **expanded uncertainty**, U_c. Note: k is typically in the range 2 to 3 (ISO, p. 3; Fluke, p. 20-6).
(e.g., If the combined standard uncertainty is $u_c = 0.3$ and a coverage factor of $k = 2$ is used, then the expanded uncertainty is $U_c = ku_c = 0.6$)

law of propagation of uncertainty — the uncertainty δ_z of a quantity $z = f(w_1, w_2, \ldots, w_N)$ that depends on N input quantities $w_1, w_2, \ldots, w_N$ is found from

$$\delta_z^2 = \sum_{i=1}^{N} \left(\frac{\partial f}{\partial w_i} \right)^2 \delta_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\partial f}{\partial w_i} \frac{\partial f}{\partial w_j} \delta_i \delta_j \delta_{ij}$$

where δ_i^2 is the variance of w_i and δ_{ij} is the correlation coefficient of w_i and w_j. If the input quantities are independent (as is often the case), then the correlation is zero and the second term of the above equation vanishes. The above equation is traditionally called the "general law of error propagation," but this equation actually shows how the uncertainties (not the errors) of the input quantities combine (ISO, p. 46; Bevington, p. 43; Taylor, p. 75).

Example: $V = \pi r^2 h$, with $r = 5.0 \pm 0.1\,\text{cm}$, and $h = 10.0 \pm 0.3\,\text{cm}$

$$V = \pi (5.0\,\text{cm})^2 (10.0\,\text{cm}) = 785\,\text{cm}^3$$

$$\delta_V^2 = \left(\frac{\partial V}{\partial r} \right)^2 \delta_r^2 + \left(\frac{\partial V}{\partial h} \right)^2 \delta_h^2 + 2 \frac{\partial V}{\partial r} \frac{\partial V}{\partial h} \delta_r \delta_h \delta_{rh}$$

where $\dfrac{\partial V}{\partial r} = 2\pi hr$, $\dfrac{\partial V}{\partial h} = \pi r^2$, and $\delta_{rh} = 0$ if r and h are not correlated

$$\delta_V^2 = (2\pi hr)^2 \delta_r^2 + \left(\pi r^2\right)^2 \delta_h^2$$

$$\delta_V^2 = [2\pi (10\,\text{cm})(5\,\text{cm})]^2 (0.1\,\text{cm})^2 + [\pi (5\,\text{cm})^2]^2 (0.3\,\text{cm})^2$$

$$\delta_V^2 = 993\,\text{cm}^6 + 555\,\text{cm}^6$$

$$\delta_V = 39.3\,\text{cm}^3 \quad \text{and} \quad \frac{\delta_V}{V} = \frac{39}{785} = 5.0\%$$

$$V = 785 \pm 39\,\text{cm}^3 \quad \text{or} \quad 790 \pm 40\,\text{cm}^3 \quad \text{when properly rounded.}$$

Note: In this example, the absolute uncertainty in h is larger than for r, but because the radius is squared, δ_r contributes nearly twice as much as δ_h to the total uncertainty in V.

Alternative approach:

The above calculation can be simplified by dividing both sides of the equation by V^2 to yield an equation in terms of relative uncertainties:

$$\left(\frac{\delta_V}{V} \right)^2 = \left(2 \frac{\delta_r}{r} \right)^2 + \left(\frac{\delta_h}{h} \right)^2 = (2(2\%))^2 + (3\%)^2 = (4\%)^2 + (3\%)^2$$

$$\frac{\delta_V}{V} = 5\% \qquad \text{(same relative uncertainty as above)}$$

2.2 Reporting Uncertainties

When reporting the measurement of a physical quantity, some quantitative estimate of the quality of the result should be given so that people who use the result can assess its reliability. Without such an indication, measurement results cannot be compared, either among themselves or with theoretical or reference values. Unfortunately, many scientists and engineers do not explicitly report the uncertainty of their measurements, so that the reader is forced to assume that the result is known to the precision implied by the number of significant figures. For example, $v = 20.2$ m/s implies an uncertainty of ± 0.1 m/s or ± 0.5%. However, there are many cases where data are improperly reported with excessive precision (extra digits) that is not justified by the experimental procedure, a practice that is careless, misleading, and could even be considered unethical.

Even when the uncertainty in a measured value is explicitly reported (e.g., ± 0.1 m/s), the meaning is not always clear because there are various methods and formats for reporting uncertainties. The following table shows the most common formats:

Table 2-1. Common formats for reporting uncertainties

Example	Explanation	Reference
$m = 2.32$ g with a combined standard uncertainty $u_c = 0.05$ g	u_c is the combination of all Type A (statistical) and Type B (systematic/other) errors; denotes	ISO *Guide to the Expression of Uncertainty in Measurement.*, 1993.

	approx. a 68% confidence level.	
m = 2.32 g with an expanded uncertainty U = 0.10 g	Calibration certificates usually report a 95% confidence level with coverage factor k = 2.	*NIST Calibration Services Users Guide 1998*, p. 4.
m = 2.32 ± 0.05 g	The meaning of ± 0.05 is vague and depends on various conditions; "reasonably certain" measured quantity lies in this range; margin of error.	J. Taylor. *Error Analysis*, 1997 p. 14.
m = (2.32 ± 0.05) g	The uncertainty generally represents ± 1 or the 68% confidence level for the measurement.	P. Bevington & K. Robinson. *Data Reduction and Error Analysis for the Physical Sciences*, 1992, p. 39.
m = 2.32 ± 0.10 g	In the field of chemistry, the uncertainty generally represents the 95% confidence level.	
m = 2.324(52) g	"numbers in parentheses indicate experimental uncertainties in last two digits" This notation is common in atomic and nuclear physics.	Table of fundamental constants found in several popular physics textbooks. E. R. Cohen, B. N. Taylor, *Rev. Mod. Phys.* 1987, 59:1121.
accuracy = ± (1% of reading + 2 digits)	Manufacturers typically specify instrument tolerance limits, which generally represent a 99% confidence level, but may be 95% or some other confidence level depending on marketing strategy.	Fluke. *Calibration: Philosophy and Practice*, 1994, p. 20-7, 22-4. Phone conversation with Fluke application engineer, Mar.1999.
m = 2.32 g ± 2% or m = 2.32 (2%) g	2% is a relative uncertainty, but the confidence level is not clear	
m = 2.32 SE 0.01 g	SE = standard error	C. David. *J. Chem. Educ.* 1996, **73**, p. 46.
55% favor candidate A (± 3% margin of error)	the margin of error in a poll generally represents a 95% confidence interval	J. Taylor. *Error Analysis*, 1997 p. 14.

As can be seen from the table above, not only are there differences in notation with essentially the same meaning, but depending on the source and context, the quoted uncertainty could represent a 68%, 95% or even a 99% confidence interval. In an effort to

avoid this kind of confusion, the International Organization of Standardization (ISO) has recently specified universal guidelines for expressing the uncertainty of measurements [ISO, 1993 #120]. These guidelines are designed to provide a uniform method for comparing measurements made in different countries in the fields of science, engineering, industry, commerce, and regulation. However, most physics teachers are not familiar with these guidelines (none of the physics instructors surveyed in this study cited the ISO Guide as a recommended reference). Consequently, students are instructed to use methods of error analysis and reporting that may not be consistent with the ISO Guide (as indicated in the table above and also in the analysis later in this chapter). Because there are various methods for treating measurement uncertainty, an important part of this dissertation research involves a careful examination of the instructional resources on this topic to better understand what introductory physics students are expected to know. These findings are presented in the following sections.

2.3 Summary of References

As a first step in discerning what students are expected to know about measurement uncertainty, a ranking analysis was conducted to ascertain which references are most often cited by other sources or recommended by instructors. The analysis consisted of a cross-referencing matrix created in an electronic spreadsheet to sort references according to how frequently they are cited in the bibliography section of 8 reference books and 7 journal articles on the subject of error analysis. References recommended by 10 physics instructors from the Expert Survey (Appendix D) were also included in this analysis. The following table summarizes the results:

Table 2-2. Error Analysis References, Ranked by Number of Citations

(# Cited is the number of citations made by 25 different sources.)

Author: *Title*	Years Published	# Cited
J. Taylor: *An Introduction to Error Analysis*	1982, 97	9
P. Bevington: *Data Reduction and Error Analysis*	1969, 92	8
D. Baird: *Experimentation*	1962, 95	5
Y. Beers: *Introduction to the Theory of Error*	1957	4
N. Barford: *Experimental Measurements*	1967	2
G. Box: *Statistics for Experimenters*	1978	2
H. Braddick: *The Physics of the Experimental Method*	1956	1
W. Deming: *Statistical Adjustment of Data*	1944, 84	1
C. Dietrich: *Uncertainty, Calibration and Probability*	1991	1
W. Dixon: *Introduction to Statistical Analysis*	1969, 83	1
W. Fuller: *Measurement Error Models*	1987	1
ISO: *Guide to the Expression of Uncertainty in Measurement*	1993	1
W. Lichten: *Data and Error Analysis*	1988, 99	1
L. Lyons: *A Practical Guide to Data Analysis for Physical Science Students*	1991	1
J. Mandel: *The Statistical Analysis of Experimental Data*	1964, 84	1
H. Margenau: *The Mathematics of Physics and Chemistry*	1943, 47, 56	1
S. Meyer: *Data Analysis for Scientists and Engineers*	1975	1
M. Natrella: *Experimental Statistics*	1963, 66, 83	1
F. Pugh: *The Analysis of Physical Measurements*	1966	1
B. Schigolev: *Mathematical Analysis of Observations*	1965	1
G. Squires: *Practical Physics*	1965, 85	1
C. Swartz: *Used Math*	1993	1
NIST/B.Taylor: *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results* http://physics.nist.gov/Pubs/guidelines/contents.html	1994	1
E. Wilson: *An Introduction to Scientific Research*	1952	1
A. Worthing: *Treatment of Experimental Data*	1946	1

H. Young: *Statistical Treatment of Experimental Data*	1962	1

The books by Taylor and Bevington appear to be the most popular references, and each is cited by about one third of the resources surveyed. Taylor's book provides a basic introduction to error analysis, while Bevington's book covers the topic at a higher level suitable for upper-division undergraduate or graduate students. Baird's book, which is similar to Taylor's, is the only other current publication that is frequently cited. Interestingly, the *Guide to the Expression of Uncertainty in Measurement,* published by the International Organization for Standardization (ISO) in 1993, was not cited by any of the physics experts or reference books. (Its one citation came from the NIST guidelines, which were adapted from the ISO Guide). It is surprising that the ISO Guide is not referenced more often, because this document is now recognized by industry as the primary reference on this subject. It could be argued that the ISO Guide is not cited frequently because it is a relatively new publication. However, there does not appear to be a strong correlation between the age of a reference and the number of citations in the above table since the Pearson correlation coefficient between these variables is only $r = 0.2$ for the top 10 references. More specifically, Taylor' book is cited most frequently despite the fact that it was first published after many of the less popular books. Based on conversations with physics teachers and graduate students, it appears that the ISO Guide is simply not well known in academia. In fact, in a phone conversation with the author of the NIST guidelines, Dr. Barry Taylor encouraged me to help "spread the word" about the ISO Guide methods to the American Association of Physics Teachers (AAPT), the American Physical Society (APS), and the American Chemical Society (ACS) [Taylor, 1999 #269]. He said that over 30,000 free copies of the Guide have been requested and distributed to users, and a modified

version of the Guide is available to the public on the NIST website, but evidently the Guide is still not widely known and used in the physics community. The consequences of this lack of familiarity are apparent from the expert responses to questions related to the uncertainty of measurements, as discussed in Chapter 4.

2.4 Previous Studies on Students' Understanding of Measurements

As noted earlier, very little research has been documented to assess introductory physics students' understanding about measurement uncertainty. In fact, searches in the ERIC database revealed only two published studies that explore university physics students' conceptions about measurement errors and the reliability of experimental data. (The Educational Resources Information Center, ERIC, is the largest database for education research.) Two additional papers were discovered through cross-references and direct contact with the authors. These other two papers examined middle-school children's understanding of measurements. It is interesting to note that none of these studies were conducted in the United States, and all four studies examined students in different countries: France, Great Britain, South Africa, and Korea. Numerous other instructional references on measurements and error analysis were found (see Table 2-2), but none of these addressed the epistemologies of the learner.

In 1993, Sere, et al. analyzed students' concepts about the need for repeated measurements, distinctions between random and systematic errors, and their notion of confidence intervals (Sere, Journeaux et al. 1993). This study involved detailed examination of the second-semester laboratory work of twenty first-year physics students at the University of Paris. From observations and follow-up interviews, the researchers learned

that most of the students did not understand the significance of confidence intervals as demonstrated by their failure to consider the uncertainty of their measurement when deciding whether their measurements were consistent with each other. The researchers were surprised that none of the students drew graphical representations of their results to examine the global view of the measurements. The students were also generally reluctant to take more than one or two measurements to find the focal length of a lens, and when asked to make a series of ten measurements, they often placed more confidence in their first measurement and used subsequent measurements to judge the preceding ones. Despite prior instruction on measurement errors and the use of statistics to analyze multiple measurements, these students failed to recognize the purpose of taking repeated measurements. Students also confused systematic and random sources of error, and the concepts of precision and accuracy were also not clearly distinguished by many students. As Thomson (1997) points out, this terminology is not used consistently even in physics publications.

Lubben and Millar (1996) surveyed over 1000 United Kingdom students aged 11, 14, and 16 about the reason for repeating measurements, how to handle repeated measurements and anomalous readings, and the significance of the spread in a set of data. They identified a pattern of progression in the understanding of empirical data with age and experience (see Table 2-3). They also suggested that other research tools using interviews should be developed for further investigation into students' conceptions about measuring, accuracy and precision, random and systematic errors, sample size, and the evaluation of small differences between measurements to decide if the difference is significant or not.

Table 2-3. Model of progression of ideas concerning experimental data

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

Note: Levels A-H were proposed by Lubben and Millar, while category I was proposed by Allie et al.

The suggestions made by Lubben and Millar were pursued by a group of researchers who conducted a study in 1998 to examine 121 first-semester physics students and their ideas about the reliability of experimental data [Allie, 1998 #217]. This study at the University of Cape Town, South Africa, used written questions and interviews with students to confirm many of the findings of Lubben and Millar and extend their model of ideas concerning experimental data (Level I in Table 2-3). Even though the students in this study were older than those in the secondary school study, the model proposed by Lubben and Millar was still useful for classifying the procedural ideas of these university students who mostly fell into levels F, G, and H. The study used nine written "probes" or scenarios all related to the same experimental situation where a ball is released from rest, rolls down a ramp, and lands on the floor some distance *d* from the edge of the table on which the ramp is secured. Findings from six of the probes are presented in the paper (there is no mention of the remaining three probes). Three of the probes dealt with the reasons for repeating

measurements and the other three dealt with sets of experimental data (how to handle an anomalous measurement, how to compare two sets of measurements having the same mean but different spreads, and how to compare two sets of measurements having similar spread but different means). A clear majority (58%) of the students reasoned that measurements of the distance and time the ball fell needed to be repeated in order to establish an accurate mean value. The remaining students were classified into nearly even clusters of thinking. One cluster (7%) did not see a purpose in repeating *distance* measurements, but all of these "non-repeaters" reasoned that several *time* measurements need to be taken. Another small cluster (8%) of "repeaters" believed that additional time and distance measurements are needed to practice and perfect the experimental process of taking measurements. The final cluster (10%) of "confirmers" suggested repeating distance measurements in order to find a recurring value. Responses to the probes that dealt with sets of experimental data showed that students are not able to differentiate clearly between the overall spread of the data set and the differences between the individual data points within the set. Details about the findings from these data set probes can be found in Chapter 4, where similar questions were examined for this study.

A 1998 study (unpublished) conducted in Korea investigated the measuring abilities and conceptions of thirty middle-school students (age 14) (Soh, Fairbrother et al. 1998). These students were asked to make measurements of length, time, volume, mass, and force using typical laboratory instruments. Students' ability to make accurate measurements (within the precision of the measuring instrument) ranged from 4% to 97% depending on the task. Details about several of these measuring tasks are presented in Chapter 4. The students were also interviewed about their conceptions on repeated measurement, use of

several measurements of the same quantity, and measurement uncertainty. The researchers found that a majority of the students repeated measurements only if they felt that their earlier measurements were inaccurate. Students were asked "Do you think completely accurate measuring is possible? If it is possible, how can you achieve it?" To this question, 41% of the students answered affirmatively, stating that tools or machines like computers could give accurate measurements, but humans can not unless they are trained well. Only one student said that both man and machine can make errors. These results indicate that about half of the students do not understand the inherent nature of uncertainty in measurements.

In summary, these previous studies addressed several of the broader issues related to measurements: the reasons for repeated measurements, concepts about accuracy and precision, random versus systematic errors, the treatment of anomalous data, and assessing the quality of measured data by the mean and spread. However, none of these earlier studies examined the process by which students determine and quantify the uncertainty of a measurement, which is the focus of this dissertation study. In all of the above studies, the measurements made by students were analyzed on their own merit and without comparison to measurements made by instructors or other "experts." The studies generally failed to indicate the level of uncertainty that students should be expected to achieve. This omission will be examined in substantial detail in this study where student responses to measurement questions will be compared to responses given by instructors and other "experts" who are familiar with these issues. Relevant components of each of these earlier studies have been incorporated into the design of this dissertation study, and whenever possible, comparisons are made between the current and previous findings.

2.5 What Are Students Expected to Know and Practice?

After examining the existing references on this topic, the next step in this research project was to organize a list of student learning objectives pertaining to measurement uncertainty. An initial list of 50 objectives were generated from personal experience conducting laboratory experiments and from teaching other physics students. This list was refined by examining the major reference books on error analysis. A focus group with eight physics education researchers was conducted to further examine what other physicists feel are the key issues that should be addressed by this study. After analyzing the focus group discussion, the list of learning objectives was revised. This revised list was presented to a group of about 25 physics graduate students and professors who rated each learning objective on a scale of 1 (lowest) to 5 (highest) for three different criteria:

1. How important is this concept for introductory physics students to understand?
2. How well do introductory physics students understand this concept?
3. How well do you personally understand this concept?

This survey can be found in Appendix B, "Learning Objectives Survey." The results of this survey, combined with responses from the Expert Survey (n = 28), are shown below.

Expert responses to the question:

> *What do you think are the most important concepts or skills students should learn about measurement uncertainty and error analysis?*

Note: The following statements were written by experts, and were edited only enough to clarify meaning. The bold category headings were added after the statements were compiled and sorted. This procedure is consistent with the "grounded theory" approach to qualitative research, where theoretical models are allowed to emerge from the empirical data. (Strauss and Corbin 1990)

All measured values have uncertainty

Every measurement has uncertainty no matter how careful you are.
All measurements have a certain level of unavoidable uncertainty.
All physical measurements have uncertainties associated with them.

All measurements are uncertain.
All measured values have uncertainty.
Every measurement has some kind of uncertainty associated with it.
Uncertainty results from estimation using tools of known precision.
All measurement tools have limits.
Know the accuracy of your apparatus.
Physical quantities are never exactly known (like π or *e*).
Not all results have a "theoretical value." The value quoted in textbooks is usually an "experimental value."
Always present
Measurement results are not exact, but are in a range of results governed by a distribution law. There are different types of probability distributions, and we often use the normal distribution.

Uncertainties must be estimated and clearly reported

We must clearly convey the size of uncertainties to our readers.
How the uncertainty is reported must be stated and must match the type of data and the needs of your audience.
How a number is to be used determines how it is usually reported.
The necessity of providing measurement uncertainties.
Importance of accurate estimation and reporting of uncertainty.
Reporting uncertainties (acknowledge your ignorance!)
Estimate and report random errors.
Uncertainty should be reported in labs (via sig. figs. or other method).
How to correctly present results from labs.
How to *estimate* an error.

Reporting proper number of significant figures

Meaning of sig. figs./uncertainty.
Significant figures – students are often insanely precise for one measurement when others are very imprecise.
How to meaningfully interpret the results of a computer calculation: (i.e., all 14 places are *not* significant).
Not to report all digits on the calculator (i.e., significant figures).
Truncate measured values according to the order of possible error.

Propagation of errors

Know how to propagate uncertainty.
Methods exist for determining the uncertainty in a computed result (propagation of errors) or in a slope or intercept from a graph.
Uncertainties propagate through the various calculations that are done with raw data.
Error propagation.
Estimate uncertainty in calculated numbers from uncertainty in data.
How to calculate uncertainty.

Identify and classify sources of error

Be able to identify and classify sources of error in data.
Difference between systematic and random errors.
Systematic vs. random errors, selection effects.
What a systematic error is.
How to differentiate between human error and systematic error.
What, why, and where certain kinds of errors occur.
Types of error: random, systematic, etc.
Relative source of errors.
Sources of uncertainty.
Where does the error occur in the setup, the equipment?
Distinctions among different kinds of uncertainties (imprecision, inaccuracy, limits of resolution, etc.)

Interpreting and reducing errors

Physical interpretation is the error low or high? What does that say about the experiment, and what should I do about it?
Taking multiple measurements reduces random error, but does not reduce systematic error.
How to reduce errors.
When human error is negligible in comparison with other errors.

Use of uncertainty for comparing results or designing experiments

Think about the uncertainties when comparing different estimates for the same value.
Comparing results requires knowledge of the uncertainties.
The most important issue for me is that students understand the function of error analysis i.e., that the results of experiments are the subject of discourse in communities of scientists, and that statistical measures can serve to constrain this discourse. For example, the community may not accept a claim unless it can be demonstrated that it is statistically significant at $p < .05$; it could even specify what sorts of tests should be done, e.g., chi-squared. Viewed this way, error analysis should be part of experimental design and the execution of experiments, and not something that you do *after* the experiment. Of course, *defending* experimental results in debate rarely is part of what students do, at least in introductory courses.
Error analysis is connected to experimental design, and this allows us to compare two different experimental designs with the same aim (e.g., using one photogate or two to measure the acceleration of a cart on an incline).

Other

Skeptical attitudes towards dogmas about uncertainties (e.g., “a result is worthless unless you quote an error,” “you must always put error bars on a graph”)
The difference between a theoretical and experimental value.
The meaning of "confidence interval."

Don't discard data unless it is the result of instrument malfunction or your own mistake. (This is a serious problem in industry.)
How to linearize functions (Linear regression analysis will most likely be used at some point in their careers.)
Understand what one should expect in a problem.
There are no right answers, but there are wrong answers.
Use of error bars on graphs.

Based on the above responses, it appears that the expert respondents believe it is most important for introductory physics students to understand the fundamental principles of measurement uncertainty, and that proficiency in performing detailed error analysis is not as important.

2.5.1 How Did Experts Learn Error Analysis?

As part of the Expert Survey (Appendix D) that was administered to physics graduate students and professors, the following sources were listed in response to the question: "Where or how did you learn to analyze errors in measurements?"

Table 2-4. Ways experts learned error analysis

How experts learned error analysis:	#Cited
Undergraduate lab classes and manuals	12
Teaching and performing lab experiments	6
Physics classes	5
Statistics courses	3
College chemistry class	2
Sophomore year in college	2
On the job training and practice	2
Journal articles	1
Calculator manuals	1
Books	1
Experimental nuclear physics	1
Astrophysics	1

High school math class	1
Graduate school advisor	1
Common sense	1

Clearly the most significant way that physics experts learn error analysis is from studying or teaching undergraduate lab classes. Therefore, it is prudent to ensure that the error analysis instruction presented to students in introductory physics lab courses is accurate.

2.5.2 Expert and Novice Approaches to Physics Problem Solving

Since this research compares students' ability to analyze measurement errors with that of experts, it is worthwhile to present previous research findings on general differences related to how experts and novices solve physics problems. A large number of research studies have been conducted to examine expert and novice differences in problem solving, and so only the most relevant points will be summarized here.

The approach to problem solving is different for novices and experts. Experts work forward toward a solution while novices generally attempt a working-backward approach using a means-ends analysis (comparing what is given in the problem with what is to be solved and trying to reconcile the difference) (Larkin 1981) (Chi, Glaser et al. 1983). Novices tend to classify and solve problems according to the surface characteristics (e.g., a spring problem) instead of the underlying physics principles (e.g., conservation of energy) (Chi, Feltovich et al. 1981). Expert problem solvers usually draw a picture or diagram to help them think about the problem, while novices often skip this step and jump immediately to analyzing the problem using quantitative equations. Experts add this qualitative analysis step to better understand the problem from a broader perspective (Larkin and Reif, 1979; Chi, Glaser, and Rees, 1983).

2.6 Summary

Because of the various conventions that are used to discuss the uncertainty of measurements, it has been necessary to first identify and clarify these conventions by conducting a thorough review of the reference literature on error analysis. These reference materials and expert survey responses have identified the student learning objectives that should be addressed by this research.

3 Research Procedures

3.1 Research Methodology

This study primarily uses qualitative research methods to gain a deeper understanding of students' epistemologies about measurement uncertainty. Quantitative analyses of differences within and between groups of students are performed in cases where sufficiently large sample sizes allow for meaningful differences to be observed. Several different types of research procedures are used to obtain insights into students' understanding from a variety of perspectives in an effort to triangulate upon a more accurate and balanced view rather than one that may be biased by examining only from a single perspective. These research procedures include:

1. a focus group and survey with colleagues to study the key issues that should be addressed in this study,
2. a written student survey on measurement uncertainty designed to address the key issues,
3. follow-up interviews with students to clarify their responses to the survey,
4. a written expert survey designed to compare and contrast differences between expert and novice responses,
5. an analysis of student laboratory reports, quizzes, and homework assignments to get an authentic perspective of how students communicate their ideas in their coursework,
6. interviews with students on laboratory procedures and experiment design,
7. and a lab practicum for students and experts to demonstrate their procedural knowledge in obtaining physical measurements.

3.2 Qualitative Analysis

Even though the subject of this study is numerical in nature (since the uncertainty in measurements can be quantified), *qualitative* research methods are primarily used to examine students' treatment of uncertainty. An inductive *grounded theory* approach has

been taken with this research to allow patterns and constructs to emerge from dense empirical data. This approach is described by Strauss and Corbin (1990) as distinctly different from the more traditional scientific process of formulating a hypothesis that is then tested against empirical observations. The reason for the grounded theory approach is that this research is formative in nature, so the methodology should be broad-based and not confine the extent of the empirical data. However, even an open-ended investigation must have some direction in order to reach meaningful conclusions that address the questions that motivated the study. The direction for this research is guided by feedback from physics instructors as described later in this chapter.

The qualitative research methods used in this study are based on practices suggested by Gall, Borg, and Gall (1996), Miles and Huberman (1994), and Strauss and Corbin (1990). As is common in most qualitative research, the data were *coded* into themes or categories based on patterns observed through repeated words, phrases, or numerical data that emerged from the student responses. The coding process was often revised and repeated as additional data necessitated the modification of existing categories. The accumulation and analysis of data was terminated when *saturation* was reached (no new findings emerged from additional data) or the available pool of data was exhausted. A computer spreadsheet program (Excel) was the primary research tool used to organize and analyze the research data gathered for this study. This program proved to be a flexible and effective tool that facilitated both qualitative and quantitative analysis of the data.

3.3 Quantitative Analysis

Although a variety of data sources were used to examine how students treat uncertainty in measurements, the same quantitative data analysis procedures are used throughout this study and consist of descriptive statistics and hypothesis testing. Since the qualitative data is categorized, the fraction of students in similar categories can be compared across sample groups to determine if there is a significant difference between the sample proportions. A *z*-test can then be used to examine the difference between sample means. Since the proportions are dichotomous (a student response is either in a category or not in that category), the sampling distribution for each proportion is defined by a binomial distribution. The uncertainty associated with a proportion p from a sample size n is the standard deviation of the binomial sampling distribution:

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

The sampling distribution for the sample proportion p is approximately normal if $np \geq 10$ and $n(1-p) \geq 10$ (Moore 1995). A reference table with uncertainty values calculated for key proportions p and alternative proportions $q = 1 - p$ is shown below.

Table 3-1. Relative uncertainty values for binomial distribution

	Probability Splits (p%)/(q%)				
n	**10/90**	**20/80**	**30/70**	**40/60**	**50/50**
5	13.4%	17.9%	20.5%	21.9%	22.4%
10	9.5%	12.6%	14.5%	15.5%	15.8%
20	6.7%	8.9%	10.2%	11.0%	11.2%
30	5.5%	7.3%	8.4%	8.9%	9.1%
50	4.2%	5.7%	6.5%	6.9%	7.1%
100	3.0%	4.0%	4.6%	4.9%	5.0%
200	2.1%	2.8%	3.2%	3.5%	3.5%
500	1.3%	1.8%	2.0%	2.2%	2.2%
1000	0.9%	1.3%	1.4%	1.5%	1.6%

5000	0.4%	0.6%	0.6%	0.7%	0.7%
10000	0.3%	0.4%	0.5%	0.5%	0.5%

From this table, we can see that the uncertainty values decrease as the sample size increases and also as the proportions move away from a 50/50 split. For example, a sample with n = 100 and a 50% proportion has a standard deviation of 5%, while the uncertainty for a 10% proportion is 3%. Since most of the sample sizes in this study are less than 100, the proportions of student responses have uncertainty values that are at least 3%. This means that these proportions should be rounded to two significant figures so that the excessive precision is not implied by (in)significant digits (see section 4.5.1 on relative uncertainty and significant figures). Consequently, all proportions tabulated in the results section of this report are rounded to the nearest whole percentage point.

Just as the students in this study should consider the uncertainty of their measurements when designing an experiment, this same consideration is necessary for this research about the students. Since comparisons will be made between groups of students, the expected variation in the response rates for each group should be used to determine the minimum sample size needed to show a meaningful difference between the groups. The required sample size could be estimated from the binomial uncertainty table above, but a better procedure is to consider the hypothesis test that will be used to examine the difference between proportion means.

If the sample sizes are sufficiently large (see below), then a z-test can be used to compare the population proportions, p_1 and p_2 by considering the null hypothesis that these proportions are equal: Ho: $p_1 = p_2$. The alternative hypothesis is that these proportions are different: Ha: $p_1 \neq p_2$. The test statistic is then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

where $\hat{p}_1$ is the proportion of responses in the category of interest for one group and $\hat{p}_2$ is the proportion of responses in the same category for the other group. The pooled proportion estimate, $\hat{p}$, is the weighted average of the combined sample proportions and is given by

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

The *z*-test statistic represents the ratio of the difference in proportions to the standard deviation of the distribution of proportion differences. This statistic can be used for comparing proportions if for each sample, more than five observations fall in the category for which the proportion is estimated, and more than five observations do not fall in that category. If this condition is not satisfied, then the distribution of the test statistic will not be sufficiently close to normal, and the *Fisher's exact test* should be used (Agresti and Finlay 1997). Since the Fisher's exact test is accurate for both small and large sample sizes, it will be used for all analyses. An example of the SAS procedure and data output for this analysis can be found in the Appendix.

The probability associated with a test statistic is found from the standard normal distribution. The *p*-value is the two-tailed probability found from a normal distribution table or a calculated value from a computer program. If the *p*-value is below a specified significance level ($\alpha = 0.05$ for this study), then the null hypothesis is rejected, and there is sufficient evidence to accept the alternative hypothesis that the sample proportions are different. With an established significance level of $\alpha = 0.05$, there is only a 5% chance that

a *p*-value less than 0.05 will result in an incorrect decision to accept the alternative hypothesis, when in fact the null hypothesis is true (this is called a Type I error). A Type II error would occur if the null hypothesis were accepted when there truly was a significant difference between the sample proportions. These same judgment issues arise when students compare their experimental values with a predicted value, each of which have some degree of uncertainty (see section 4.7.1).

The minimum sample size required to show a meaningful difference between two sample proportions can be found from the above equations. If we believe that a proportion difference of at least 0.2 is meaningful, then sample sizes of at least fifty will yield a statistically significant difference between proportions p_1 and p_2. Somewhat smaller sample sizes could yield this same 20% resolution between p_1 and p_2 if the proportions are far from 50%. For example, if both sample groups have 20 students, then a significant difference (p-value < 0.05) can be observed for $p_1 = 0\%$ and $p_2 = 20\%$. These minimum sample sizes were considered during the design phase of this study, and a sample size of 50 students was set as a target value.

3.4 Determining the Key Issues to Investigate

In addition to a careful review of the error analysis references already mentioned, several other methods were employed to narrow the scope of this research and obtain feedback from expert practitioners who have direct experience with students dealing with measurement uncertainty issues. While the reference books provide a comprehensive view of error analysis, they generally do not indicate the areas that provide the greatest conceptual difficulties for students. In order to investigate this cognitive aspect, it was necessary to

consult directly with instructors who teach about measurements, and with the students themselves. The following sections describe how this was accomplished.

3.4.1 Review of Topics in Reference Books

As discussed in Chapter 2, a concerted effort was made to determine what students are expected to know about the uncertainty of measurements. For this reason, a sample of the most popular reference books on the subject, along with physics textbooks, laboratory manuals, and other error analysis guides were reviewed to gain further understanding of the instructional content. This review provided a context for addressing the conceptual understanding of students and the difficulties they encounter in analyzing the uncertainty of measurements.

3.4.2 Focus Group

A focus group was convened in November, 1997 to get feedback on an initial list of perceived areas of difficulty that students encounter related to measurement uncertainty. A transcription of the focus group, including the questions that were addressed, is provided in Appendix K. This meeting included eleven members of the North Carolina State University Physics Education Research (PER) group – seven graduate students, three professors, and one administrator. The hour-long discussion was recorded on audiotape, and then transcribed. The comments expressed in this focus group helped shape the direction of this research study, as explained in Section 2.5.

3.4.3 Survey of Learning Objectives

A survey (Appendix B) was developed to solicit feedback from experts on the primary learning objectives that they felt should be addressed by this research. This survey was administered in the early stages of this research to about 25 physics graduate students and professors at North Carolina State University. The participants discussed the issues in groups of three or four and submitted their notes after the hour-long meeting. A summary of these findings has already been presented in Section 2.5.

3.5 Measurement Uncertainty Survey

One of the primary research instruments for this study was a written survey with open-ended questions to address the objectives recommended by the experts. The questions for this survey were designed to cover a broad range of objectives while still having some overlap between questions to provide reliability checks. The survey was designed to require less than thirty minutes for a typical student to complete, but students actually spent anywhere from 10 to 60 minutes to complete these surveys. Student volunteers were interviewed to determine if the questions were sufficiently clear and whether they elicited the desired responses. Throughout the development phase, the survey was revised several times based on discussions with colleagues and responses from student interviews. The final product consisted of two versions (A and B) to allow for greater breadth of topics that could be examined while keeping each survey to a reasonable length. These surveys are provided in Appendix C.

3.6 Expert Survey on Measurement Uncertainty

A second survey for experts (Appendix D) was developed to gather responses from instructors to establish a "standard" to which student responses to similar questions could be compared. The experts who completed this survey included physics professors and graduate teaching assistants. Invitations to experts were publicized on the physlrnr listserve, meetings of the American Association of Physics Teachers (AAPT), and the North Carolina Section of the AAPT. An effort was also made to solicit responses from chemistry professors and graduate students; however, only three of these people returned their surveys, so nearly all of the 28 experts who completed this survey are from the physics community.

Approximately twenty experts outside of academia were also contacted by telephone or in person and questioned about their practices of determining and reporting uncertainties in measurements. These experts included industrial metrologists, application engineers, calibration engineers, and NIST employees. While many of these conversations were helpful in understanding calibration and control processes, most were not directly relevant to this study. When asked specific questions about the expression of uncertainty, most of these industrial contacts referred to the methods presented in the ISO Guide. References to conversations with these experts have been included in this study when appropriate.

3.7 Population and Sample Description

The target population for this study is introductory physics students, which includes high school, college, and university students taking an introductory physic course. The primary subjects in this study were all university students in either a first or second-semester physics course. For comparison purposes, graduate teaching assistants (TAs) assigned to

these courses were also included in the study. A majority of the research findings come from data gathered at North Carolina State University (NCSU), but data were also obtained from the University of North Carolina at Chapel Hill (UNC) and the University of Hokkaido in Japan (Hokudai). The North Carolina universities were primarily chosen for their accessibility to the researchers, but they are also believed to be representative of typical universities, which is an important consideration when generalizing findings beyond the scope of the sample being investigated.

3.7.1 NCSU Sample Description

Nearly all of the physics students from NCSU who participated in this study were engineering majors. They were enrolled in the first or second-semester calculus-based physics courses (PY205 and PY208). Both of these courses used the textbook by Halliday, Resnick, and Walker: *Fundamentals of Physics*, 5th ed. Multiple sections of these large-enrollment courses follow a similar curriculum since the students take common exams that are administered simultaneously across campus. Each of these 4-credit hour courses includes a required laboratory component, which counts for 10% of the students' course grade. Students meet for lab every-other week and perform six laboratory experiments throughout the semester. A complete, written laboratory report is required for each experiment. The laboratory curriculum is typical of many university physics labs, and about half of the experiments utilize personal computers for data acquisition and analysis.

Nearly all of the NCSU students in this study are sophomores (65%) or juniors (25%) who have already taken a chemistry course, which includes a laboratory component where many error analysis concepts are introduced. The proportion of female students in this

NCSU sample ranged from 15% to 39%, which is comparable to the overall proportion for the university (40%), and the fraction in the school of engineering (19%).

Table 3-2. NCSU student sample demographics

Course #	Course Description	Course Content	Lab Practic.	Sample Size	Female Fraction
PY205	Calculus physics for engineers	Mechanics – waves	Survey	28	15%
PY208	Calculus physics for engineers	E & M – modern	Survey	71	18%
PY205	Calculus physics for engineers	Mechanics – waves	Ver. A	37	27%
PY205	Calculus physics for engineers	Mechanics – waves	Ver. B	36	39%
PY208	Calculus physics for engineers	E & M – modern	Ver. A	34	32%
PY208	Calculus physics for engineers	E & M – modern	Ver. B	32	31%

The Student Measurement Uncertainty Survey was administered to nearly 100 students during the last laboratory period of the fall semester in 1998. The administration of this survey was conducted in conjunction with pre-post testing to evaluate the gains made by students in their conceptual understanding of course-specific physics topics. Students who participated in the pre and post testing received extra credit equivalent to a 100% on one additional laboratory report. A representative sample of laboratory sections was selected to collect data from at least thirty students in each of the two courses taught by six different instructors. Students in these sections were asked to complete the measurement uncertainty survey instead of one of the post-testing instruments, and they received the same extra credit as their peers.

3.7.2 UNC Sample Description

The UNC students in this study were enrolled in one of four classes. Physics 24 and 25 are the required introductory physics courses for pre-medical students and other health science majors. This algebra-based sequence used the textbook by Serway and Faughn,

College Physics, 5th ed. Physics 26 and 27 are the first two semesters of physics for students who plan to major in physics, chemistry, computer science, or other technical majors that require calculus-based physics. This sequence used the textbook by Halliday, Resnick, and Walker: *Fundamentals of Physics*, 5th ed.

Each of the 4-credit hour UNC physics courses includes a required laboratory component, which counts for 25% of the student's overall course grade. Students meet for lab every week and perform nine laboratory experiments throughout the semester. A complete, written laboratory report is required for each experiment. The laboratory curriculum is typical of many university physics labs and is similar between the algebra and calculus-based tracks. Therefore, both of the first semester students (Phys24 and 26) were given the same lab practicum for mechanics experiments (Appendix F.1), and the second-semester students (Phys25 and 27) used the same activity designed for electricity and magnetism experiments (Appendix F.2).

The UNC lab practicum was administered at the end of the fall semester in 2000 as a makeup lab activity for students who had missed a lab sometime during the semester. These students were assessed on their performance and received scores that were normalized to 85% (the average lab score) and counted the same as a regular laboratory report score. The grades were assigned by the students' regular laboratory instructor and were based on the grading rubric that was designed for this activity.

Two significant differences exist between the laboratory curriculum at UNC compared with NCSU. The first difference is that the UNC labs do not use computers to collect data with interface probes like the NCSU labs do. Students only use computers to analyze their data using KaleidaGraph or Excel software. The second and most important difference for

this study is that the topic of error analysis is emphasized much more in the UNC physics labs than it is at NCSU. Students are required to estimate the uncertainty in their results (calculating standard errors and propagating uncertainties as needed) for practically every UNC physics experiment. This same level of rigor is not emphasized in the NCSU labs. This difference between the two curricula is apparent when comparing the level of detail in the measurement sections of each lab manual (Appendix A.1 and A.2). One of the reasons that samples were selected from each of these two schools was to examine the hypothesis that students from the UNC sample should demonstrate a better understanding of measurement uncertainty than students from NCSU.

Table 3-3. UNC student sample demographics

Course Number	Course Description	Course Content	Lab Exam	Sample Size	Female Fraction
Phys24	Algebra physics for pre-meds	Mechanics – waves	I	23	52%
Phys25	Algebra physics for pre-meds	E & M – nuclear	II	14	64%
Phys26	Calculus physics for scientists	Mechanics – waves	I	17	47%
Phys27	Calculus physics for scientists	E & M – Optics	II	9	44%

The total enrollment at UNC is similar to NCSU, but the admissions standards are higher for UNC (see average high school Grade Point Average and SAT scores in Table 3-4). About 50% of the UNC students in this study are female, which is slightly lower than the overall university enrollment which is 60% female.

Table 3-4. NCSU and UNC student population statistics

	NCSU 1997	NCSU 1999	UNC 1997	UNC 1999
Total enrollment	27,529	28,011	24,189	24,635
Undergraduate	19,097	19,027	15,321	15,434
Women	40.1%	41.1%	60.1%	60.6%
White	81.3%	81.0%	81.3%	81.2%

African American	12.1%	10.6%	11.1%	11.2%
Freshmen Profile:				
Average SAT Verbal	567	577	609	620
Average SAT Math	587	602	611	625
Average SAT Combined	1154	1179	1220	1245
H.S. GPA	3.69	3.86	4.02	4.06

Sources: www2.acs.ncsu.edu:80/UPA, www.ais.unc.edu/ir

3.7.3 Hokudai Sample Description

In the summer of 1998, I had the opportunity to travel to Japan and Korea to collaborate with other physicists who were also interested in students' understanding of measurements. In Japan, I collaborated with Dr. Shigeo Sugiyama, a professor of science history in the physics department of Hokkaido University (also called Hokudai). In South Korea, I met with Jongah Soh, a physics graduate student at Seoul National University, who under the direction of Dr. Sungjae Park, was investigating how well her junior high school students understood measurements they made. While these collaborations were not the primary focus of my dissertation study, the discussions and insights that resulted from these visits were invaluable.

In addition to collaborating with Japanese and Korean researchers, I also planned to compare the responses of Japanese and Korean students with answers from similar American students. My expectation was that the Asian students would demonstrate a higher level of understanding about measurement uncertainty than their American counterparts. This hypothesis was based on the superior past performance of these groups of students on the Third International Mathematics and Science Study (TIMSS, 1996).

Table 3-5. TIMSS rankings for selected countries

Country	Math rank (score*)	Science rank (score*)
Singapore	#1 (643)	#1 (607)
Japan	#3 (605)	#3 (571)
South Korea	#2 (607)	#4 (565)
England	#25 (506)	#10 (552)
United States	#28 (500)	#17 (534)
South Africa	#41 (354)	#41 (326)

*Average score of 13-year olds on TIMSS.
Average score of all 41 countries = 500

Time and resources did not permit a detailed investigation of Korean students; however, I was successful in gathering responses from over 150 Japanese students at Hokkaido University, thanks to the gracious assistance of Dr. Sugiyama, who coordinated the administration of my Measurement Uncertainty Survey in several different classes.

Hokkaido University has a total enrollment of about 12,000 undergraduate students and 5,000 graduate students. Of the approximately 200 universities in Japan, Hokudai ranks in the top 10. Like NCSU, Hokudai was originally an agricultural college. The college of medicine is now its largest "faculty."

The Japanese school year begins in April, so these students were in their first semester when they were surveyed in June, 1998. An experimental physics class is required of all physics majors, and most take this lab course after completing their first year of physics. Since all of the students who participated in this study were in their first year, none had taken this lab course, but nearly all had physics laboratory experience from high school.

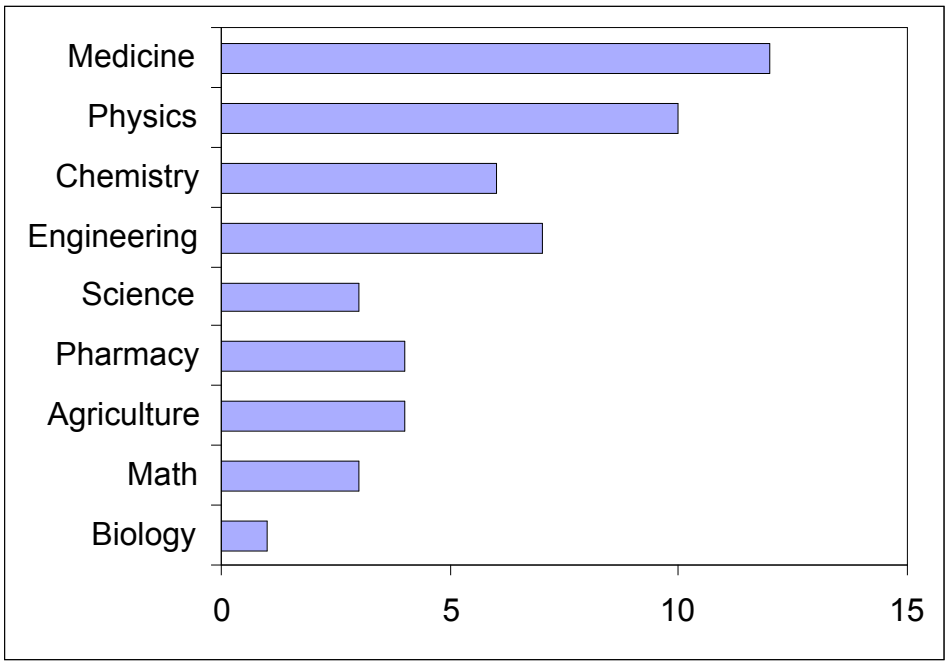


Figure 3-1. Hokudai sample: number of students represented in each major

The Measurement Uncertainty Survey (Appendix C) was translated into Japanese by Dr. Sugiyama (Appendix C.3 and C.4). Both the English and the Japanese versions of the survey were delivered to students in six different courses: history of science, English, biology, physics, chemistry, and engineering. The students were not given any compensation for completing and returning the survey, so the only motivation was obligation (their instructor asked them to do this) or kindness. The procedure for delivery and collection of the surveys was not tightly controlled since Dr. Sugiyama was also acting as coordinator on a volunteer basis. Consequently, the response rate varied from 100% in classes where students were asked to complete the survey in class (i.e., History of Science, taught by Dr. Sugiyama) to only about 20% in the chemistry and engineering classes. This low response rate is a serious threat to the internal validity of the sample because the

students who did respond were essentially self-selected. It is believed that self-selecting students would feel more confident in their ability, and therefore should perform at a higher level than a randomly-selected sample from the same population.

Follow-up interviews were conducted with 20 students to clarify their responses to the survey questions. These twenty students were selected based on their willingness to be interviewed, their availability during the interview period, and their major. In addition, two graduate students were interviewed to gain another perspective from outside the population of interest. One of Professor Sugiyama's graduate students, Kaori Takaguchi, assisted with translation during the interviews.

3.7.4 TA Sample Description

The responses of graduate teaching assistants (TAs) were included in this study to serve as a standard to which the student responses could be compared. The TAs who participated included all of the laboratory instructors for the NCSU and UNC students who were the primary subject of this investigation. These TAs are considered to be valid experts since they were the ones most familiar with the laboratory curriculum, and they were responsible for assessing the laboratory and data analysis skills of the students they taught. However, these TAs are not considered to be authorities on measurement uncertainty since their experience and training is primarily limited to their physics teaching experience and training, research lab experience, and their own undergraduate laboratory experiences as a student. The TAs participated primarily out of obligation as part of their job responsibilities, so their motivation was also different than that of the students. This difference in motivation is a potential threat to the validity of the responses.

3.8 Physics Lab Practicum

An important research tool used in this study was a laboratory practicum that was developed to assess students' procedural knowledge. The Lab Practicum tested students' ability to make accurate measurements, correctly use common laboratory equipment, and analyze experimental data. The questions on this exam were selected to cover the topics and types of activities required in the lab course, with approximately equal numbers of direct measurement and computational questions. The exam was administered at the end of the semester both at NCSU and UNC. The NCSU students received extra credit for their effort equivalent to 100% on one additional lab grade. The UNC students took the practicum as a makeup lab activity, so unlike their NCSU counterparts, their performance was assessed and their score (normalized to 85%) counted toward their course grade.

3.8.1 Interviews on Experimental Design

Six experimental design interviews were conducted during the development phase of the Lab Practicum. Two or three students participated in each interview as they explored an open-ended investigation with questions designed to address key aspects of each experiment (Appendix H). The students in these interviews were from a special section of the NCSU PY208 course, and they received extra credit towards their lab grade for volunteering to participate. Each student signed an informed consent form that explained the objective, procedure, and potential consequences of the research (Appendix G). These experimental design interviews were used primarily to guide the development of the Lab Practicum, so a detailed analysis of the student procedures and responses was not performed.

3.9 Limitations

As with any qualitative research, there are limitations to the ability to generalize the research findings to the broader population of interest, and even more limitations to consider when practitioners try to apply the findings to their own situation. While the student samples selected for this study were chosen to be representative of typical university physics students, the findings from this study may not be consistent with all groups of students within this target population. The most significant internal and external threats to validity are presented below.

3.9.1 Threats to Internal Validity

Threats to internal validity are factors that can confound an observed difference between an experiment group and a control group. The internal validity of an experiment is the extent to which extraneous variables have been controlled by the researcher, so that any observed effect can be attributed solely to the treatment variable (Gall, Borg et al. 1996). This study did not employ a traditional experimental design to examine the effect of varying one single variable and observing the outcome, but comparisons are still made between different sample groups, and the factors that could obscure any observed differences between these groups should be considered.

- **Differential selection**

The observed differences between responses from student groups could depend as much or more on the inherent differences between these groups (e.g. intelligence, prior education, and experiences of the students) as on the differences in the physics laboratory curriculum. Random assignment of students into treatment and control groups is the best

safeguard against differential selection, but this was not possible (or even relevant) since this study was not a traditional experiment design. However, it is expected that the effect of these individual differences would be less noticeable with increased sample size.

- **History**

Observed differences between groups of students may be affected by other events or factors that occur over a period of time. This study examines different groups of students with varying degrees of experience with physics. It is quite possible that factors other than their physics instruction could influence the ability of these students to analyze measurement problems. The most significant of these additional factors are knowledge and experience gained from courses in statistics, chemistry, and laboratory research experience.

3.9.2 Threats to External Validity

Threats to external validity are factors that limit the ability to generalize the findings of a study. External validity is the extent to which the findings of a study can be applied to individuals and settings beyond those that were studied (Gall, Borg et al. 1996).

- **Population validity**

There is inherent risk in generalizing from the sample of students selected from the locally accessible population of students at NCSU and UNC to the larger target population of introductory physics students nationwide.

- **Ecological validity**

The generalization of findings from this study are also limited by the extent to which the environmental conditions of the study approximate the actual conditions of the subjects in the target population. For this study, these concerns relate primarily to whether or not the

research instruments are “authentic” or contrived. The most authentic sources of data for this study come from the analysis of student lab reports and homework. These regular student assignments provide a natural source of data, unlike the student surveys, interviews, and lab practica, which were developed or conducted specifically for this research. These research instruments are subject to a variety of factors which can influence student performance.

- **Motivation**

The students in this study were generally volunteers who received extra credit or sometimes no tangible reward for their willingness to answer questions. These same students might give different answers if given the same questions as a graded homework assignment or on an exam where the stakes are higher.

4 Research Findings

4.1 Overview

Detailed findings from each of the research methods described in Chapter 3 are presented here, organized by subject, from simple to complex levels of reasoning. The order follows the list of topics that experts believe students should know (Section 2.5), and this same order is preserved in Chapter 5, where a taxonomy of student difficulties is presented as a summary of these findings. The vast majority of data from this research comes from the Lab Practicum and the Measurement Uncertainty Survey, as these two research tools proved to be the most useful in gathering the desired breadth and depth of insight into students' treatment of measurement uncertainty. Findings from student interviews, lab reports, and homework supplement these primary research instruments.

4.2 The Nature of Uncertainty in Measurements

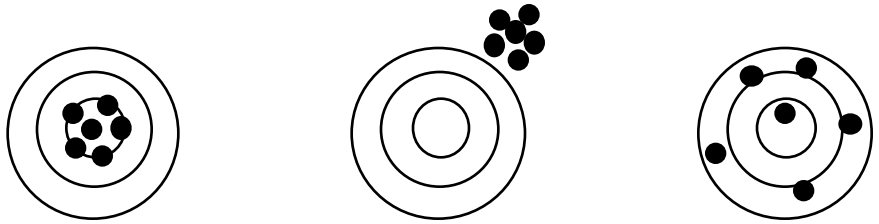
All measurements have some degree of uncertainty, no matter how carefully the measurement was obtained. However, a significant number of students (~50%) believe that exact measurements can be made if high-quality equipment is used and there is no "human error" (mistakes made by the person taking the measurement) (Soh, Fairbrother et al. 1998).

4.3 Accuracy, Precision and the Use of Standards

Students often confuse precision and accuracy. A common mistake made by students is to assume that a precise instrument or measurement is also accurate, when this may not be the case (i.e., there is a systematic error). Misconceptions about measurements

that may be "precisely wrong" suggest a lack of understanding for the need to calibrate equipment referenced to traceable standards.

One example of the confusion between these terms was explored by asking NCSU students to rate the accuracy and precision of the target shooting scenarios depicted in Figure 4-1. A tally of their responses is provided in Table 4-1.



Accuracy: Good
Precision: Good

Accuracy: Poor
Precision: Good

Accuracy: Good?
Precision: Poor

Figure 4-1. Accuracy versus precision – target shooting example
Reference: (Doran 1980)

Table 4-1. Student rating of precision and accuracy

n = 61		A	B	C
Accuracy	Good	**59 (97%)**	2 (3%)	**13 (21%)**
	Poor	2 (3%)	**59 (97%)**	48 (79%)
Precision	Good	**59 (97%)**	**48 (79%)**	3 (5%)
	Poor	2 (3%)	13 (21%)	**58 (95%)**

Bold indicates "correct" answers according to the author of this study.

Student responses are consistent with the definitions of the terms accuracy and precision, except for target C, where the shots are scattered (low precision), but on average they are centered on target (good accuracy). One explanation is that students were asked to evaluate the accuracy and precision of diagrams similar to, but not identical to the one above. In one case, the last target was drawn with the shots scattered and none were within

the inner circle or "bullseye". Of the 31 students in this particular group, 29 (94%) rated the accuracy low. This response rate was significantly different ($p = 0.009$) for another group of students in the same course (same population) where 19/29 (66%) of the students rated the accuracy low. In this second class, one of the shots was drawn within the bullseye. In discussing the responses with students, several of the students said that they rated the accuracy high for target C because of the mark in the center. This finding is similar to when students get 0% error and do not consider the uncertainty of their experimental result (they ignore the scatter).

4.4 Reporting the Best Estimate of a Measured Value

Before examining students' practices related to the uncertainty in a measurement, we should first discuss the proper procedure for determining and reporting the best estimate of the measured value itself. This procedure is quite simple if only one measurement is made: the measured value should be reported to a reasonable number of significant digits, along with a variable name and appropriate units. (Ex. Diameter = 3.25 cm)

When multiple measurements (replicates) are made of the same value, then the sample *mean* (average) is most commonly used to represent the central tendency of the data set. The mean for n individual measurements is defined as: $\bar{x} = \frac{\sum x_i}{n}$. If the sample distribution is skewed by extreme values, and there is sufficient justification to omit these *outliers* from the data set (see below), then the mean of the remaining values may be used to represent the best average value. If there is no good reason to omit outliers from the data set, then the

median (middle value) is considered to be a better estimate of a typical central value for the sample (Moore 1995).

4.4.1 Recognizing Anomalous Data

A common problem that arises when making measurements and analyzing data is the question of how to treat anomalous data. Most experts agree that data should not be discarded without good reason, but what criteria should be used to decide whether an outlier should or should not be omitted from the data analysis? The simplest and safest solution is to never discard any measurements. However, this practice of including outliers may significantly skew the sample data set so that the mean is not the best estimate of the target value. The solution that generally yields the most accurate results is to apply Chaevenet's criterion, which states that a data point should be discarded if less than half an event is expected to lie further from the mean than the suspect measurement (Bevington and Robinson 1991; Taylor 1997). This criterion accounts for outliers that will exist with some predictable probability depending on the sample size and variance. Another reasonable criterion is to discard a data point if it lies more than 3 standard deviations away from the mean, but to do so with reasonable judgment, especially if the data set is small (Baird 1995). Perhaps the best solution is to re-examine the suspect data point and repeat the measurement if possible. Many great scientific discoveries have been made from investigating what first appeared to be an anomaly.

In the South Africa study (Allie, 1998), one of the probes presented two alternatives for dealing with an anomaly, and students were asked to choose which one they agreed with and explain their reasoning:

A group of students have to calculate the average of their (distance) measurements after taking six readings. Their results are as follows (mm): 443, 422, 436, 588, 437, 429.
The students discuss what to write down for the average of the readings.
A: "All we need to do is to add all our measurements and then divide by 6."
B: "No. We should ignore 588 mm, then add the rest and divide by 5."

Table 4-2. Summary of responses to the South African anomaly probe

Category Description	Frequency of response $n = 121$
The anomaly must be included when taking an average since all readings *must* be used	37 (30%)
The anomaly is noted, but it has to be included in the average since it is part of the spread of results	14 (12%)
The anomaly must be excluded as it is most likely a mistake	30 (25%)
The anomaly must be excluded as it is outside the acceptable range	38 (31%)
Not codeable	2 (2%)

Only about half of the students excluded the anomaly, which lies about 19 standard deviations from the mean of the other 5 data points. Such a distant outlier is almost certainly a mistake and should be excluded from the data analysis. It is surprising that more students did not exclude the anomalous point, especially when the students were explicitly confronted with the question of whether or not the suspect data should be omitted from the average.

In an effort to replicate the above findings and investigate how readily students recognize an outlier, the South Africa probe was modified and included in the Student Measurement Uncertainty Survey (Appendix C) administered to NCSU and Hokudai students. The Lab Practicum (Appendix F) administered to NCSU and UNC students also included these questions. Version A asked the following question:

A group of students are told to use a meter stick to find the length of a hallway. They take 6 independent measurements (in cm) as follows: 440.2, 421.7, 434.5, 523.4, 437.2, 428.9. What result should they report? Explain your answer.

Version B asked the same question with one different data point (492.5 instead of 523.4):

440.2, 421.7, 434.5, 492.5, 437.2, 428.9

The data could be graphically represented on a number line (not shown to students):

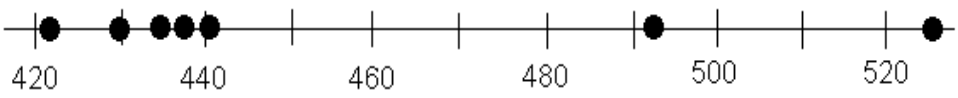


Figure 4-2. Data points for length of hallway problem
(Note that the anomalous data points lie well beyond the cluster of other measurements.)

The purpose for the two different versions of the same question was to see if there was any difference in how students treated the anomalous data point depending on whether it was 1.0 m or 0.5 m from the mean of the other 5 measurements. The measurements 523.4 and 492.5 were also chosen to see if students more readily recognized an outlier when the first digit is different from the other measurements. In both cases, the suspect measurement lies at least 8 standard deviations from the mean, and therefore is most likely a mistaken result (possibly a counting error from repeatedly moving the meter stick). The data sets were also chosen so that the average of either the 5 or 6 data points yielded a mean ending in an even digit followed by a 5. The purpose of this unusual condition was to see if students would round their result up or down in accordance with two different recommended procedures (Bevington and Robinson 1991; Serway and Beichner 2000).

According to their laboratory instruction manuals (Appendix A), students should answer this question by omitting the extreme outlier and calculating the mean and standard

error of the remaining five measurements: $L = (432 \pm 3)$ cm. However, most of the NCSU and UNC students averaged all six data points and reported the mean to four significant figures with no uncertainty estimate. Here is a summary of how students answered this question:

Table 4-3. Summary of responses to treatment of data question

Student treatment of data	Hokudai Students $n = 52$	NCSU-A Students $n = 36$	NCSU-B Students $n = 37$	NCSU TAs $n = 13$	UNC Students $n = 40$	UNC TAs $n = 10$
Number who reported a value	30	29	27	11	37	10
Averaged all 6 data points	15%*	62%*	89%*	27%	80%	80%
Omitted single outlier and averaged other measurements	**35%**	**17%**	**4%**	**64%**	**8%**	**20%**
Omitted high and low from avg.	50%	3%	4%	0	0	0
Reported median or other value	4%	17%	3%	9%	0	0
Reported 4+ significant figures	86%	87%	96%	73%	86%	30%
Reported 2 or 3 significant figs.	**14%**	**13%**	**4%**	**27%**	**14%**	**70%**
Showed explicit uncertainty	**10%**	**10%**	**11%**	**82%**	**73%**	**100%**

Bold indicates response that is most correct according to expert opinion.
* indicates significant differences between samples

Significant differences were observed among the sample groups in how they treated the data and reported a best estimate. A much higher fraction (85%) of the Hokudai students omitted one or more data points before calculating an average value, while only about 15% of the NCSU students, and only 8% of the UNC students rejected the outlier before finding a mean value. This wide discrepancy in the treatment of the data prompted further investigation. During the follow-up interviews, several of the Japanese who had omitted both the highest and lowest value explained that they had learned this "trimmed mean" procedure in their statistics class, and they gave an example from the Olympics of dropping the high and low scores for skating competitions. Except for the Hokudai students, the percentage of students and TAs who excluded the outlier was significantly less than the ~50% rate of the South

African students who were explicitly confronted with this issue. It appears then that most students (and TAs) followed the routine practice of calculating an average value without considering the distribution of the data. This conclusion is supported by the lack of drawings similar to Figure 4-2 on any of the student (and only 2 of the TA) papers. It is quite likely that a larger fraction of the students would have recognized and rejected the outlier if they had plotted the data points on a number line to visualize the distribution of the data. As stated by Box (1978, p. 25), "a dot diagram is a valuable device for displaying the distribution of a small body of data (up to about 20 observations)."

Slight differences were also observed between the two versions of this data analysis question. As expected, a larger fraction (17%) of NCSU students recognized and omitted the outlier 523.4 from the data set compared with the 4% fraction of students who omitted the outlier 492.5. However, these differences are not statistically significant at the 0.05 level, so no general conclusions can be drawn. The seemingly large difference between the proportions of NCSU and UNC TAs who averaged all 6 measurements may also be due to random chance since the difference is barely significant at the 0.05 level (the 2-tailed probability from Fisher's exact test is $p = 0.03$).

4.4.2 Ability to Make Accurate Measurements

Several questions on the Lab Practicum assessed students' ability to make simple measurements. Two such questions asked students to measure the diameter of a penny as accurately as possible with both a ruler (1 mm resolution) and with a Vernier caliper (0.05 mm resolution). The ability of students and TAs to make and report an accurate measurement was surprisingly low for such a simple task. Only about 60% of NCSU

students and 70% of the TAs accurately reported the diameter of penny measured with a ruler within 0.5 mm, and 15% did not correctly give a value within the 1 mm precision of the ruler. The UNC students demonstrated a significantly higher level of competence with 88% of the students reporting accurate values within 0.5 mm and 100% within 1 mm of 1.90 cm, which is the median value of all responses. It is surprising that a higher percentage of the UNC students performed better on this task than did their teaching assistants, but the difference is not statistically significant ($p = 0.065$). The one UNC TA whose answer was not within the 1 mm resolution reported a diameter of 0.12 cm, which is clearly a mistake.

The most significant difference between the student groups was for the proportion of students who reported an explicit uncertainty value (as required). Even though nearly all of the students remembered to include proper units with their measurement, *none* of the 37 NCSU students reported an uncertainty value, while about 40% of the UNC students did so. Section 4.5 discusses this issue of reporting uncertainties in greater detail.

Table 4-4. Measuring the diameter of a penny with a ruler (1 mm resolution)

	NCSU		**UNC-CH**		**Signif.**
D = 1.90 ± 0.05 cm	**Students**	**TAs**	**Students**	**TAs**	**Diff.?**
	***n* = 37**	***n* = 6**	***n* = 40**	***n* = 10**	**(Fisher)**
within 0.5 mm (1.85 to 1.95 cm)	21 (57%)	4 (67%)	35 (88%)	6 (60%)	**0.0042**
within 1 mm (1.8 to 2.0 cm)	31 (84%)	5 (83%)	40 (100%)	9 (90%)	**0.0098**
value reported to 1 sig. fig.	4 (11%)	0 (0%)	0 (0%)	0 (0%)	**0.049**
value reported to 2 sig. figs.	26 (70%)	5 (83%)	26 (65%)	4 (40%)	0.64
value reported to 3 sig. figs.	7 (19%)	1 (17%)	14 (35%)	6 (60%)	0.13
explicit uncertainty reported?	0 (0%)	2 (33%)	18 (45%)	7 (70%)	**<0.001**
reasonable uncert. (0.025 to 0.1 cm)	0 (0%)	1 (17%)	15 (38%)	6 (60%)	**<0.001**
units shown?	35 (95%)	6 (100%)	40 (100%)	10 (100%)	0.23
wrong units? (do not match value)	8 (22%)	2 (33%)	4 (10%)	1 (10%)	0.21

The significant difference column gives the probability for the 2-tailed Fisher's exact test that compares the proportions of NCSU and UNC students; p-values less than 0.05 are shown in bold to indicate statistically significant differences between these proportions.

The performance for all groups dropped considerably when using the Vernier calipers to obtain the same measurement within the same error limits. Ironically, it appears that students lose *accuracy* with this measuring instrument that allows for much greater *precision*. What is even more surprising is that only two-thirds of the students and TAs in both groups reported diameter measurements that were consistent with each other. The failure of one-third of students to recognize a discrepancy between their two measurements of the same quantity with different instruments is another indication that students do not associate a meaning to the quantities that they measure and report – they are just numbers.

Table 4-5. Measuring the diameter of a penny with calipers (0.05 mm resolution)

	NCSU		UNC-CH		Signif.
D = 1.905 ± 0.005 cm	**Students**	**TAs**	**Students**	**TAs**	**Diff.?**
	***n* = 37**	***n* = 6**	***n* = 40**	***n* = 10**	**(Fisher)**
within 0.5 mm (1.85 to 1.95 cm)	20 (54%)	5 (83%)	27 (68%)	7 (70%)	0.25
within 1.0 mm (1.80 to 2.00 cm)	21 (57%)	6 (100%)	30 (75%)	8 (80%)	0.10
value reported to 1 sig. fig.	4 (11%)	1 (17%)	0 (0%)	0 (0%)	**0.049**
value reported to 2 sig. figs.	22 (59%)	2 (67%)	6 (15%)	1 (10%)	**<0.001**
value reported to 3 sig. figs.	7 (19%)	2 (67%)	21 (53%)	2 (20%)	**0.0040**
value reported to 4+ sig. figs.	2 (5%)	1 (17%)	13 (33%)	7 (70%)	**0.0034**
explicit uncertainty reported?	0 (0%)	2 (33%)	15 (38%)	6 (60%)	**<0.001**
reasonable uncert. (0.002 to 0.05 cm)	0 (0%)	1 (17%)	11 (28%)	5 (50%)	**<0.001**
units shown?	27 (73%)	6 (100%)	39 (98%)	10 (100%)	**0.0026**
wrong units? (do not match value)	5 (14%)	2 (33%)	4 (10%)	0 (0%)	0.73
agrees with ruler measurement?	24 (65%)	4 (67%)	29 (73%)	7 (70%)	0.62

The significant difference column gives the probability for the 2-tailed Fisher's exact test that compares the proportions of NCSU and UNC students; p-values less than 0.05 are shown in bold to indicate statistically significant differences between these proportions.

A third question on the Lab Practicum asked students to determine the radius of a steel ball as accurately as possible using any available equipment. The intent of this question was to see how many students would correctly use the Vernier calipers (instead of a ruler) without explicit instructions to do so. The student performance on this task was

disappointingly low, with only about half the students reporting an accurate radius measurement. A summary of the results are shown in

Table 4-6 Accurately finding radius of a steel ball using any available equipment

	NCSU		UNC-CH		Signif.
R = 0.79 ± 0.01 cm (NCSU)	**Students**	**TAs**	**Students**	**TAs**	**Diff.?**
R = 0.951 ± 0.001 cm (UNC)	***n* = 36**	***n* = 7**	***n* = 40**	***n* = 10**	**(Fisher)**
Radius reported within 1 SD	**5 (14%)**	**3 (43%)**	**18 (45%)**	**6 (60%)**	**0.0054**
Radius reported within 5 SD	10 (28%)	4 (57%)	22 (55%)	6 (60%)	**0.021**
Diameter reported instead of R	7 (19%)	2 (29%)	6 (15%)	4 (40%)	0.76
Reported using calipers	**13 (36%)**	**0 (0%)**	**24 (60%)**	**7 (70%)**	**0.043**
used calipers correctly	**2 (6%)**	**0 (0%)**	**18 (45%)**	**7 (70%)**	**<0.001**
Value reported to 1 sig. fig.*	15 (42%)	2 (29%)	1 (3%)	0 (0%)	**<0.001**
Value reported to 2 sig. figs.*	**15 (42%)**	**3 (43%)**	18 (45%)	2 (20%)	0.82
Value reported to 3 sig. figs.*	5 (14%)	2 (29%)	**20 (50%)**	**8 (80%)**	**0.0013**
Value reported to 4+ sig. figs.*	0 (0%)	0 (0%)	1 (3%)	0 (0%)	1.0
Explicit uncertainty reported	1 (3%)	3 (43%)	17 (43%)	6 (60%)	**<0.001**
Reasonable uncertainty	**1 (3%)**	**1 (14%)**	**11 (28%)**	**5 (50%)**	**0.0027**
Units shown	31 (86%)	7 (100%)	35 (88%)	10 (100%)	1.0
units consistent with value	**31** (86%)	**7** (100%)	**33** (83%)	**10** (100%)	0.76

* Number of significant figures reported has been adjusted to account for those who reported diameter instead of radius.
The significant difference column gives the probability for the 2-tailed Fisher's exact test that compares the proportions of NCSU and UNC students; p-values less than 0.05 are shown in bold to indicate statistically significant differences between these proportions. "Correct" responses are also shown in bold.

The most significant difference between the NCSU and UNC groups is the percentage of students who reported an explicit uncertainty value. Although these particular measurement questions for the penny and sphere did not explicitly ask students to include an estimate of their uncertainty, this requirement was clearly stated at the beginning of the instruction sheets. Despite this, *only one* of the NCSU students included an uncertainty estimate on any of these three questions, while about a third to half of the UNC students did so. Of the UNC students who did report an explicit uncertainty, about three-fourths had

reasonable uncertainty values (as defined by expert judgment). This issue of determining and reporting uncertainty values is explored in greater detail in the following section.

The TAs from both schools reported uncertainty values with their measurements more readily than the corresponding groups of students, but the UNC TAs consistently reported uncertainty values more frequently (and correctly) than the TAs from NCSU. This difference reflects the level of emphasis placed on error analysis in the laboratory curricula of the two schools.

4.5 Determining and Reporting the Uncertainty of a Measurement

The ability of students to estimate the uncertainty of a measurement is a skill that is necessary for the higher-level task of evaluating the quality of the measurement. Unfortunately, many students fail to report an uncertainty for a measurement, even when they are explicitly asked to do so. Of those who do report a quantitative estimate of the uncertainty, this value often indicates a level of precision that is unreasonably low or high. One instructor emphasized this point on the Expert Survey: "My philosophy is that a value for the uncertainty is necessary, but the mathematics should be kept as simple as possible. The point is that the result should be *reasonable*."

The following tables summarize how UNC students reported uncertainty values on the Lab Practicum. (Less than 10% of the NCSU students reported explicit uncertainty values for these lab practicum questions, so a meaningful analysis of this small sample group was not possible with such a low response rate.) Some of the questions explicitly or implicitly required an uncertainty value as noted in **Error! Reference source not found.**. Questions labeled "no mention" mean that students were not reminded to include an uncertainty value.

However, the general instructions on the first page of the practicum stated that "for questions that require a numerical result, write your answer as you would for a formal lab report or scientific journal to indicate an appropriate degree of accuracy (proper number of significant figures and uncertainty)."

Table 4-7 Student reporting of uncertainty values from UNC lab practicum

Question from Lab Practicum	Type*	Uncertainty Required?	# of Sig. Figs				Response Frequency
			1	2	3	4	
1. Measure radius of steel ball	M	no mention	17				17 (43%)
2. Report length of hallway	C	no mention	6	6	17	1	30 (75%)
4. Report sin(85° ± 1°)	C	suggested	26	7	2		35 (88%)
5. Meas. dia. of penny with ruler	M	no mention	18				18 (45%)
6. Meas. dia. of penny with caliper	M	no mention	15				15 (38%)
7. Find race car accel. from graph	C/M	required	6	8	1		15 (38%)
9. Report accel. of falling ball	C	required	15	11	6	2	34 (85%)
10. Find rotating mass from data	C	suggested	8		1		9 (23%)
11. Meas. density of nickel coin	M	no mention	5	1			6 (15%)
12. Meas. g with pendulum	M	no mention	2	3			5 (13%)
		Average:	**12**	**3.6**	**2.7**	**0.3**	**18 (45%)**

*M/C indicates whether the question required a direct measurement, calculations, or both.

The response rate for reporting explicit uncertainty values ranged from 13% to 88% on these questions, with lower response frequencies corresponding to questions that required direct measurements and did not explicitly remind students to include an uncertainty estimate. The response frequency also appears to have diminished with time, since about 45% of the students included uncertainty estimates with direct measurements at the beginning of the practicum, but the response frequency decreased to about 15% by the end of the practicum (which took students about two hours to complete).

Somewhat surprisingly, the UNC students reported uncertainty values with an appropriate number of significant figures (1 or 2) about 90% of the time. Students only reported uncertainty values with excessive precision (3 or more significant figures), on

questions that required a calculated result. A similar trend was observed for the *result* values (as opposed to the uncertainty values), as shown in Table 4-8.

Table 4-8. Student reporting of significant figures for UNC lab practicum

Question from Lab Practicum	Type[2]	n	1	2	3	4	5+	U/O[4]	d
6. Meas. dia. of penny with caliper	M	40	0	6	21	**12**	1	U	-0.80
4. Report sin(85° ± 1°)	C	39	4	4	**25**	**5**	1	U	-0.67
5. Measure dia. of penny with ruler	M	40	0	26	**14**	0	0	U	-0.65
1. Measure radius[1] of steel ball	M	40	1	17	**18**[3]	4	0	U	-0.48
12. Measure g with pendulum	C/M	32	0	0	**16**	16	0	O	0.50
2. Report length of hallway	C	37	0	3	**2**	32	0	O	0.78
7. Find race car accel. from graph	C/M	36	2	**11**	13	6	4	O	0.97
9. Report accel. of falling ball	C	38	0	**3**	31	3	1	O	1.05
10. Find rotating mass from data	C	32	0	**6**	15	7	4	O	1.28
11. Measure density of nickel coin	C	36	0	**4**	18	13	1	O	1.31

Notes:

1) The number of significant figures has been corrected to account for students who reported the diameter instead of the radius of the steel ball.
2) M/C indicates whether the question required a direct measurement, calculations, or both.
3) Bold indicates the "correct" number of significant figures for each question.
4) U/O indicates whether students tended to report values with too much (O = overly precise), or with too little precision (U = under-reported).

There appears to be a correlation between the type of question (measured or calculated) and the implied precision of the reported values. Students tend to report values with too many significant figures if the result comes from calculations, but insufficient precision is often reported for direct measurements. One explanation is that students underestimate the uncertainty in a single measurement because they often consider only the instrument precision (± 1/2 division or ± 1 division) and do not include other sources of error that contribute to the overall uncertainty. This was the case for 7 out of 10 lab groups in the French study in which 20 students were asked to measure the focal length of a lens, where *f* was not well-defined over a 4-mm range (Sere, Journeaux et al. 1993). Even though

these students recognized that a variety of factors contribute to the uncertainty in the focal length, only one group attempted to account for these factors in their evaluation of the uncertainty.

An attempt was made to confirm the findings from the French study by asking a similar question on the Lab Practicum given to the NCSU and UNC students. The students were told to use a light ray box (which can produce 5 parallel light rays) to measure the focal length of a diverging lens with less than 5% uncertainty in the measurement. The accepted answer (based on careful measurements and analysis of the TA responses) was $f = -12.5 \pm 0.5$ cm, which has a relative uncertainty of 4%. A summary of the responses given by the students and TAs is provided in Table 4-9.

Table 4-9. Uncertainty values reported for the focal length of a lens

Uncertainty (cm)	**NCSU Students**	**NCSU TAs**	**UNC Students**	**UNC TAs**
0.5	$n = 2/28$	$n = 3/5$	$n = 10/19$	$n = 2/6$
0.8		1		
0.5	1			
0.3				1
0.2			1	
0.24			1	
0.225			1	
0.10		1	2	1
0.05	1	1	3	
0.01			2	
Correct f value?	2 (7%)	3 (60%)	2 (10%)	3 (50%)
f < 0 ?	0	0	5 (26%)	0
units reported?	26 (93%)	5 (100%)	17 (89%)	6 (100%)

The response rate for reporting an explicit uncertainty was quite low on this question, especially given that the question directly stated that the uncertainty of the measurement

should be less than 5% (implying that an uncertainty value be reported). Only 2 out of the 28 NCSU students reported an explicit uncertainty, while about half of the UNC students and both groups of TAs showed uncertainty values. Even more surprising is that only about 10% of the students successfully reported a focal length measurement that was within 1 cm (2□, or 10%) of the accepted value, and only about half of the TAs did so. We should note that *none* of the students or TAs reported the focal length with a negative sign as required for a diverging lens. The only successful part of this question was that over 85% of the students and *all* of the TAs included units with their reported values (although 5 of the NCSU students reported mm when they meant cm). From these poor performance results, it is clear that there are more fundamental issues at stake than the more esoteric matter of correctly reporting uncertainty values.

4.5.1 Relative Uncertainty and Significant Figures

Even when an explicit uncertainty is not reported with a measured value, the number of significant figures implies a certain degree of precision. More specifically, the implied precision is based on the assumption that the last reported digit is uncertain. This uncertainty may be ± 0.5 or ± 1 last digit depending on the context (Taylor 1997).

Table 4-10. Correspondence between significant figures and relative uncertainty

Sig. Figs.	Value	Implied Uncertainty	Implied Relative Uncertainty
1	1	± 0.5 or ± 1	50% or 100%
1	9	± 0.5 or ± 1	5% or 11%
2	10	± 0.5 or ± 1	5% or 10%
2	99	± 0.5 or ± 1	0.5% or 1%
3	100	± 0.5 or ± 1	0.5% or 1%
3	999	± 0.5 or ± 1	0.05% or 0.1%

The number of significant figures reported in the uncertainty of a measurement should accurately reflect the appropriate confidence of the uncertainty estimate. The precision of the uncertainty value is limited by the lowest precision factor that contributes to the overall estimate of this value. In many cases, an uncertainty estimate can only be known with about 50% confidence, which means that this value should be reported with only one or two significant figures. Even if the uncertainty is represented by the standard deviation of the mean, a very large sample size (n > 10 000) would be needed to justify the use of more than two significant figures. This practice of reporting uncertainties to only one or two significant figures is consistent with nearly all of the error analysis sources referenced in this study.

Table 4-11. Relative uncertainty of the sample standard deviation

n	Exact	1/sqrt[2(n-1)]	Valid Sig. Figs.	Implied Uncertainty
2	76%	71%	1	10% to 100%
3	52%	50%	1	10% to 100%
4	42%	41%	1	10% to 100%
5	36%	35%	1	10% to 100%
10	24%	24%	1	10% to 100%
20	16%	16%	1	10% to 100%
30	13%	13%	1	10% to 100%
50	10%	10%	2	1% to 10%
100	7%	7%	2	1% to 10%
10000	0.7%	0.7%	3	0.1% to 1%

Source: ISO *Guide to the Expression of Uncertainty in Measurement*, 1993.

The approximate expression for the relative uncertainty of the standard deviation for a sample of size n is:

$$\frac{\sigma_{\sigma}}{\sigma} = \frac{1}{\sqrt{2(n-1)}}$$

From the figures in the above table, the approximate and exact expressions for the uncertainty of the standard deviation are essentially equivalent for $n > 5$.

4.5.2 Lab Practicum Question on Relative Uncertainty

Cognitive insights about relative uncertainty were gathered from both NCSU and UNC students and TAs who took the Lab Practicum that was administered at these schools. One particular question directly asked students to explore the connection between the number of significant figures and the relative uncertainty implied by a number. The question and responses are provided below:

> The number of significant figures reported for a measured value suggests a certain degree of precision. What is the relative uncertainty implied by the following numbers?
>
> a) 0.20 implies an uncertainty of ± _____ %
> b) 9.8 implies an uncertainty of ± _____ %
> c) 40 implies an uncertainty of ± _____ %
> d) 0.103 implies an uncertainty of ± _____ %

Table 4-12. Relative uncertainty responses for UNC and NCSU students

0.20 (%)	**9.8 (%)**	**40 (%)**	**0.103 (%)**	**Rationale**	**NCSU TAs** $n = 8$	**NCSU Stud.** $n = 23$	**UNC Tas** $n = 16$	**UNC Stud.** $n = 61$
5	**1**	**2.5**	**1**	**(± 1 last digit)/value**	**1**	**0**	**11**	**15**
5	1	0.25	1					1
5	2	2.5	1		1			
5	10	2.5	1		1			1
5	10	20	1			1		
5	10	100	0.5		1			
5	5	10	2.5			1		
0.05	0.01	0.25	0.01					1
0.05	0.2	1	0.003					1
0.05	0.01	0.03	0.01					1
0.05	0.01	0.03	0.029					1
2.5	**0.51**	**1.25**	**0.485**	**(± 1/2 last digit)/value**			1	2
2.5	0.5	1.25	0.049					1
2.5	0.5	1.2	1.5				1	

2.5	0.5	12.5	0.48		1			
25	0.51	12.5	4.85					1
25	0.51	10	4.85	:				1
25	2.56	12.5	2.43					1
25	5.1	12.5	4.85		1			
2.5	10	25	5			1		
1-10	1-10	1-100	0.1-1	sig. fig. table in lab manual			3	15
1	5	1	1					1
10	1	0.1	0.1					1
1	10	200	0.1					1
0.01	0.1	1	0.001	absolute uncert. in last digit		2		5
0.01	0.1	10	0.001			1		3
0.01	0.1	10	0.0001			1		
0.1	1	10	0.01			2		
0.01	1	10	0.001			1		
0.1	0.1	10	0.01			1		
0.01	0.1	0	0.001					1
0.1	10	1	0.1					1
0.1	1	2.5	0.9			1		
0.1	4.9	20	0.0515			1		
0.005	0.05	0.5	0.0005	half of last digit	2	1		
0.05	0.5	5	0.005	± 5 in last sig. fig.				2
0.01	0.05	0.5	0.005					1
10	10	20	5			1		
10	100	1000	0			1		
10	30	90	5			1		
20	5	2.5	5			1		
20	40	0	50			1		
20	980	4000	10.3	percent equivalent of value?				1
2	98	100	1.03					1
0.02	0.98	1	0.001			1		
20	10	0	30		1			
71	71	100	50	1/sqrt(n-1), n = # sig. figs.				1
47	58	7	100			1		
1	10	10	0.1			1		
1	1	2	1			1		

Table 4-13. Relative uncertainty responses for 1st and 2nd semester UNC lab students

0.20 (%)	**9.8 (%)**	**40 (%)**	**0.103 (%)**	**Rationale**	**UNC TAs1** *n* = 10	**UNC Stud.1** *n* = 38	**UNC TAs2** *n* = 6	**UNC Stud.2** *n* = 23
5	**1**	**2.5**	**1**	**(± 1 last digit)/value**	**5**	**10**	**6**	**5**
5	1	0.25	1			1		
5	10	2.5	1			1		
0.05	0.01	0.25	0.01			1		

0.05	0.2	1	0.003			1		
0.05	0.01	0.03	0.01					1
0.05	0.01	0.03	0.029					1
2.5	**0.51**	**1.25**	**0.485**	**(± 1/2last digit)/value**	**1**	**1**		**1**
2.5	0.5	1.25	0.049					1
2.5	0.5	1.2	1.5		1			
25	0.51	12.5	4.85			1		
25	0.51	10	4.85			1		
25	2.56	12.5	2.43			1		
0.05	0.5	5	0.005	± 5 in last sig. fig.				2
0.01	0.05	0.5	0.005					1
0.01	0.1	1	0.001	absolute uncert. in last digit		5		
0.01	0.1	10	0.001			1		2
0.01	0.1	0	0.001			1		
0.1	10	1	0.1			1		
1-10	1-10	1-100	0.1-1	sig. fig. table in lab manual	3	6		8
1	1	1	0.1			1		
1	5	1	1			1		
10	1	0.1	0.1			1		
1	10	200	0.1					1
20	980	4000	10.3	percent equivalent of value?		1		
2	98	100	1.03			1		
71	71	100	50	1/sqrt(n-1), n = # sig. figs.		1		

There were no significant differences at the $\alpha = 0.05$ level in the proportion of responses from the first and second semester UNC students. However, there were significant differences among the groups of TAs who answered this question correctly. None (0/8) of the TAs from NCSU gave correct values for the relative uncertainties, while over half (11/16) of the TAs from UNC correctly answered this question ($p = 0.004$). The difference in responses from these groups could be explained by the differing emphasis and exposure to this particular concept in the curricula at the two schools. What is most interesting is that there may also be a difference ($p = 0.09$) between the UNC TAs who taught the first and second semester lab courses. Despite the fact that the first semester lab

TAs had studied and answered this same question during their training session at the beginning of the semester, only 6 out of 10 answered this question correctly, while all (6/6) of the experienced TAs who taught the second-semester labs supplied the correct answer. Even with the small sample sizes, the difference between these proportions is significant at the $\alpha = 0.10$ level.

It is also surprising that more of the NCSU and second-semester UNC students did not correctly answer this question because most of these students had used WebAssign for submitting their physics homework assignments. WebAssign is an on-line homework delivery system that directly confronts students with the connection between significant figures and relative uncertainty since the default setting only accepts numerical answers within 1% of the internally calculated value. This 1% tolerance means that students must submit numerical answers with at least 2 or 3 significant figures. Evidently, this connection was either not well understood by these students, or the question on the Lab Practicum was confusing.

Students' failure to recognize the connection between significant figures and relative uncertainty can be understood partially from previous research that has examined difficulties students have in understanding ratios and proportions (Arons 1990). The findings from this study confirm that students have greater difficulty thinking in terms of proportions than absolute measures.

4.5.3 Propagation of Uncertainty in Calculations

The uncertainty in a calculated value depends on the uncertainties associated with each term used to compute the result. A conservative but simple method of estimating the

uncertainty in a result can be found by computing the maximum and minimum values based on the uncertainties of each term (this is sometimes known as the "max-min method"). The proper method of computing the uncertainty in a calculated result is to add the variances of the input quantities according to the propagation of uncertainty equation (see Section 2.1). This process of computing uncertainty values can be tedious and time consuming, but the calculations can often be simplified by ignoring terms that do not significantly contribute to the total uncertainty. An "error budget" can be compiled by listing each of the uncertainty factors and ranking them according to how much each contributes to the overall uncertainty (see Table 4-18). This technique facilitates identification of the primary source of uncertainty in a result; however, it is rarely performed by students or even instructors.

Based on interview results, students generally do not recognize that the "rules of significant figures" for addition and multiplication are simply a quick and easy way to estimate the precision of a calculated result from the errors that propagate from the original data. These rules can be stated as follows:

> When adding or subtracting measurements, the result should be rounded to the same number of decimal places as the number with the lowest precision (fewest decimal places).
>
> When multiplying and dividing, the number of significant figures that are reliably known in a product or quotient is the same as the smallest number of significant figures in any of the original factors.

While these rules of significant figures are an efficient means of propagating uncertainty and estimating the appropriate degree of precision in many calculations, they are not valid for mathematical functions like exponentials, logorithms, and trigonometric functions.

One of the questions on the Lab Practicum required students to find the uncertainty of a calculated result based on the uncertainties given for two independent factors:

A student performs a simple experiment to find the average acceleration of a falling object. He drops a baseball from a building and uses a string and meter stick to measure the height the ball was dropped. He uses a stopwatch to find an average time of fall for 3 trials from the same height and reports the following data:
$h = 5.25 \pm 0.15$ m, $t = 1.14 \pm 0.06$ s.

Use the equation $a = 2h/t^2$ to determine the average acceleration and its uncertainty.

Answer using propagation of uncertainty: $a = 8.1 \pm 0.6$ m/s^2
Answer using max-min method: $a = 8 \pm 1$ m/s^2

Table 4-14. Uncertainty reported for acceleration of falling ball

uncertainty value reported	**NCSU Students** $n = 22/36$	**NCSU TAs** $n = 7/7$	**UNC Students** $n = 34/40$	**UNC TAs** $n = 10/10$
< 0.02	1		3	
0.06	1		7	
0.07, 0.08	2		3	
0.11	1		3	
0.2	3			
0.3, 0.33	1	1	3	
0.5				1
0.6	**2**		**1**	**4**
0.7, 0.8	**2**	**2**	**1**	**2**
0.88, 0.9		**1**	**10**	**3**
1, 1.1	**3**	**3**		
> 1.2	4		3	
other	2			
% correct	**19%**	**86%**	**30%**	**90%**

The correct response rate of 30% for the first-semester UNC students is slightly higher (but not statistically different) from the 19% correct response rate of the first-semester NCSU students. However, both of these student groups performed at a level significantly below the ~90% correct response rate of their lab TAs. This suggests that practice and experience with propagating uncertainties does make a difference, but a single semester is not sufficient to master this skill.

4.5.4 Uncertainty in Slope and y-intercept from Linear Regression

Even though the issue of determining the slope from a linear regression was not a strong concern that emerged from the Expert Survey (this topic was only mentioned once), it is an issue that is commonly encountered in introductory physics laboratory experiments. Therefore, two questions on the Lab Practicum were designed to address this concept. For each question, students were given a set of non-linear data (distance versus time for a car that is accelerating, and voltage versus time for a charging capacitor). The students were instructed to analyze the data to find either the acceleration or the time constant. Unfortunately, fewer than 10% of the students (and only about 25% of TAs) analyzed these data sets correctly, so it was not possible to investigate how students treated the uncertainty in the slope of a linear regression fit for these problems. These exercises clearly demonstrate that students do not have the skills needed to decide how to analyze a set of data (this procedure is usually specified for students in their lab manuals – "plot a graph of velocity versus time"). While the issue of determining the uncertainty in the slope is important for drawing conclusions, it is secondary compared to the more fundamental task of finding a reasonable estimate of the intended result.

Despite the lack of student data for this topic, a brief discussion is warranted to address the determination of the uncertainty in the slope and y-intercept from a linear fit. If students graph their data by hand, the uncertainty in the slope can be estimated by drawing linear fits with the maximum and minimum slopes that appear to reasonably fit the general trend of the data. The uncertainty is then half this range in the slope value (Baird 1995). Likewise, the uncertainty in the y-intercept can also be estimated as half the range in the intercept values for the maximum and minimum slope fits. The easiest and most accurate

method for students to find the uncertainty in the slope and y-intercept is from a software program that automatically computes and reports these values. Several such programs are widely used in introductory physics labs (e.g., Graphical Analysis and Logger Pro by Vernier; Science Workshop and Data Studio by Pasco). Some data analysis programs, like Excel, report the correlation coefficient, r or r^2, instead of the standard error of the slope and y-intercept. For a linear fit of the equation, $y = a + bx$, to a set of n data points, the standard error of the slope b and the y-intercept a can be found from the correlation using the following formulas (Lichten 1999):

$$\sigma_b = b\sqrt{\frac{(1/r^2) - 1}{n - 2}} \quad \text{and} \quad \sigma_a = \sigma_b\sqrt{\frac{\sum x^2}{n}}$$

Determination of the uncertainty in a power-law fit or other non-linear least squares fits are beyond the scope of this study since these more complex procedures are not even addressed in many of the reference books on introductory error analysis.

4.6 Identifying Sources of Error

Measurement errors result from a variety of sources that include the precision and accuracy of the measuring instrument, the ability of the experimenter to read and interpret the measurement, and the uncertainty inherent in the phenomenon being measured. As instructed in the ISO *Guide*, all of the known sources of error should be included in the overall estimate of the uncertainty of a single measurement. However, as stated earlier from the French study (Sere, 1993), students tend to focus on the instrument precision when specifying the uncertainty in a measurement. Students also seem to believe that more expensive or high-tech instruments may reduce or eliminate experimental error (Soh,

Fairbrother et al. 1998). This sentiment is supported by one NCSU student in this study who reported that "there was no error in our experiment because we used a computer to collect and analyze our data."

4.6.1 Accuracy of Typical Physics Laboratory Equipment

Before discussing student views on the sources of error, it is logical to first examine the precision and accuracy that can be expected of measuring instruments that students typically encounter in an introductory physics lab. To this end, I conducted a cursory investigation of the physics laboratory equipment available to NCSU and UNC students. The precision of each instrument was based on its resolution. The accuracy was determined from technical specifications in catalogs, owners manuals, NIST calibrated standards, or a conservative estimate from the typical relative precision of the instrument. In general, these values can only be estimated to the nearest order of magnitude, because various grade instruments are available and most instruments can measure values over a range that is at least one order of magnitude.

Table 4-15. Typical uncertainty values for common physics laboratory equipment

Dimension	Instrument	Typical Precision	Typical Accuracy	Common Limiting Error Factor
time	digital stopwatch	0.01 s	1 to 10 ppm*	reaction time (~0.2 s)
time	photogate	0.001 s	0.01%	data processing
length	meter stick	0.5 to 1 mm	<0.5%	visual resolution
length	Vernier calipers	0.05 to 0.1 mm	0.1%	misreading scale
length	micrometer	0.001 mm	0.01%	failure to zero, misusing
mass	electronic balance	0.01 to 0.1 g	0.1%	calibration
mass	triple beam balance	0.1 g	0.1%	calibration
mass	brass mass sets	1 g	0.1%	calibration
volume	graduated cylinder	1 to 10 mL	1% to 5%	calibration
frequency	signal generator	3 to 4 digits	0.1 to 1%	calibration
voltage	multimeter	1 to 4 digits	1%	calibration

V, freq.	oscilloscope	2 digits	1% to 3%	visual resolution, calib.
current	multimeter	1 to 4 digits	1% to 5%	calibration
resistance	multimeter	1 to 4 digits	1% to 3%	extra resistance, calib.
capacitance	capacitance meter	1 to 3 digits	5% to 15%	calibration
inductance	LCR meter	1 to 3 digits	5% to 15%	calibration
mag. field	Hall probe	2 digits	5%	calibration

*ppm = parts per million (1 ppm = 0.0001% accuracy)

From the above table, it is clear why very few introductory physics experiments yield results with less than 1% error. The uncertainty of many of the above measurements is limited by the accuracy of the device, not the precision (resolution). This means that students are often confronted with situations where the measurement they obtain is more precise than it is accurate. A notable exception is for time measurements. Quartz crystal resonators are now widely used in most timing devices, and even though they are inexpensive, they provide measurements that are several orders of magnitude more accurate than any other common lab equipment. It is interesting then, that so many students (and reference books) mention the accuracy of a timer as a likely source of error: "Another source of error is that our stopwatch was not accurate [it ran too fast or slow]." – quote from student

4.6.2 Sources of Error Reported for Nickel Coin Experiment

One part of the Lab Practicum asked students to decide if nickel coins are made of pure nickel based on their measured density. The actual text from this exercise is provided here, along with a summary of the responses to this exercise.

> Use a Vernier caliper and a balance to measure the density of a nickel coin. Does your density value match (agree with) the density of pure nickel? (ρ_{nickel} = 8.912 g/cm^3). From your measured density, can you determine whether nickel coins are made of pure nickel? Which of your measurements contributes the most error to your measured density value?

Table 4-16. Measured density of nickel coins

Density (g/cm^3)	NCSU Students (n = 32/36)[1]	NCSU TAs (n = 7/7)	UNC Students (n = 36/40)	UNC TAs (n = 10/10)
Average	10.6	6.8	7.5	7.0
Median[2]	**7.2**	**7.1**	**7.1**	**7.2**
Maximum	70.3	7.5	48.4	9.2
Minimum	0.6	4.4	3.2	1.2
Std. Dev.	16.5	1.1	7.2	2.1

[1]Some samples had less than 100% response rate.
[2]Best estimate measured by author: 8.8 ± 0.4 g/cm^3.

A surprisingly wide range of density measurements was reported for this exercise. The outliers in the student samples especially skewed the average density values for these groups, so the median values are more representative of values typically reported. These median values are consistent (within 2% of each other) across all four sample groups, yet these median values are about 20% lower than the density of nickel coins based on their composition. The reason for this dramatic discrepancy is discussed below.

Table 4-17. Are nickel coins made of pure nickel?

Answer	Reasoning	NCSU Students (n = 27/36)	NCSU TAs (n = 6/7)	UNC Students (n = 32/40)	UNC Tas (n = 8/10)
no	none	16 (59%)	3 (50%)	26 (81%)	4 (50%)
no	density too low	4 (15%)	2 (33%)	3 (9%)	4 (50%)
no	cost	2 (7%)			
probably not	density too low		1 (17%)	2 (6%)	
maybe		1 (4%)			
not sure		4 (15%)		1 (3%)	
yes		0	0	0	0

According to the U.S. Mint, nickel coins are 25% nickel and 75% copper. So even though the best estimate of the measured density of the nickel coins (8.8 ± 0.4 g/cm^3) matches the density of pure nickel, we can not conclude that these coins are made of pure nickel because any number of combinations of metals could yield the same density (as is the case here). However, if the measured density of the coin was significantly *different* than 8.912 g/cm^3, then we could conclude that the coin was not pure nickel.

Table 4-18. Sources of error reported for measuring the density of a nickel coin

Source of Error	Actual Error Contribution	NCSU Students (n = 27/36)	NCSU TAs (n = 8/7)	UNC Students (n = 27/40)	UNC TAs (n = 7/10)
thickness (height)	**5% to 20%**	**4 (15%)***	**3 (38%)**	**16 (42%)***	**3 (43%)**
diameter (radius)	0.5% to 1%	5 (19%)	2 (25%)	6 (16%)	1 (14%)
mass	0.1% to 2%	11 (41%)*	3 (38%)	6 (16%)*	3 (43%)
volume		2 (7%)		2 (5%)	
reading caliper		1 (4%)		4 (11%)	
measurement error		2 (7%)		2 (5%)	
human error		2 (7%)		1 (3%)	
parallax				1 (3%)	

* statistically significant difference between student groups at the $\alpha = 0.05$ level.

The most popular source of error specified by the UNC students was the thickness (height) of the nickel coin. This source of error clearly contributes the most to the total uncertainty in the density calculation because of the indentations on the front and back faces of the coin. Many of these students correctly recognized and stated this fact in response to the question about the primary source of error. The NCSU students, however, stated the thickness at a significantly lower rate ($p = 0.028$) and instead primarily believed that the mass measurement contributed the most to the overall uncertainty in the density. One possible reason for this difference is that the NCSU students used a triple-beam balance to weigh the nickel coins, while the UNC students used a digital electronic balance. Even though both of these instruments had the same resolution (precision) of 0.1 g, there appears to be a perception by the students that the analog instrument is less accurate than the digital balance. A more carefully designed experiment with a randomly-assigned split sample would be needed to confirm this observation since the difference here is between sample populations.

It is rather surprising that even though a significant fraction of the students recognized that the raised surfaces on the coins are a source of error, nearly all of the students and TAs in this study failed to account for this factor in their calculation of the density. As a consequence, all but three students' density values were unreasonably low due to the inaccurate thickness measurement. This systematic error resulted in about 90% of the students and TAs concluding that nickel coins are not made of pure nickel, and *none* stating that the coins are pure nickel. While it is true that nickel coins are 25% nickel and 75% copper (according to the U.S. Mint), the average density of this nickel alloy is 8.92 g/cm^3, which is indistinguishable from the density of pure nickel (8.912 g/cm^3) (Weast 1988). Since the relative uncertainty of the measured density is at least ± 5%, it is impossible to resolve the 0.1% difference in densities with this measurement procedure. Despite the fact that nickel coins *should appear* to be made of pure nickel based on their density, only two out of the 76 students in this study stated that nickel coins might be pure nickel.

One additional observation from this analysis is that even though "human error" appeared several (3) times as a source of error, it was not nearly as popular a response as is perceived by laboratory instructors who regularly complain about students using this vague explanation in their lab reports.

4.6.3 Sources of Error from Student Laboratory Reports

Student laboratory reports were examined to determine if students could identify the primary source of error in an experiment. After only a brief period of investigation, it became obvious that this question could not be clearly answered because students generally did not identify the single most important source of error. Instead, they tend to cite a

"laundry list" of all possible factors that could contribute to the experimental uncertainty, perhaps hoping that at least one of these might be valid and satisfy the lab instructor who is grading the report. Unfortunately, many of these supposed sources of error were not relevant to the experiment or did not adequately explain the observed difference between the experimental and predicted results. Nearly all students fail to give quantitative arguments for the sources of error they list. Instead, these factors tend to be based on the students' "feel" for the experiment (as verified in student interviews). Nowhere in this entire study did a single student (or TA) provide an error budget which lists the sources of error along with a numerical estimate of each contribution to the total experimental uncertainty (as demonstrated in the ISO Guide and many NIST publications). Such detailed uncertainty analysis is not warranted for most introductory physics experiments; however, exposure to simple uncertainty budgets might be a useful tool for giving students a clearer understanding of which factors contribute the most to the total uncertainty.

4.7 Use of Uncertainty for Comparing Results

One of the most important reasons for determining the uncertainty of an experimental result is to provide a meaningful way to compare the result with other similar values. By comparing results, researchers can decide if an experimental result agrees with a theoretical prediction, or if results from similar studies are consistent with each other. While it is important to be able to compare experimental results with known uncertainties, it is not trivial to do so because of the inherent uncertainty in the measurements. Even when there is a prescribed procedure for deciding when results do or do not agree, the evaluation may not be reliable since the procedures for evaluating and reporting the uncertainties vary among experimenters (see

Table 2-1). Judging the agreement between uncertain results is also challenging because *evaluation* is the highest level of cognitive reasoning:

Bloom's Taxonomy of the Cognitive Domain (Bloom 1956)

1. Knowledge – memorization of facts, words, and symbols
2. Comprehension – understanding the meaning of knowledge
3. Application – applying concepts to various situations
4. Analysis – breaking apart complex ideas
5. Synthesis – putting individual ideas together to form a complete explanation
6. Evaluation – making decisions and judging the merits of ideas

As Benjamin Bloom asserts, reasoning at the higher cognitive levels (analysis, synthesis, and evaluation) requires an understanding at the lower levels. This hierarchy can explain why students struggle to make valid conclusions when evaluating empirical data. If they do not have the skills and experience necessary to comprehend and analyze their results, then the process of evaluation is nearly impossible.

The research for this section was driven by the following questions:

1. How does the evaluation process of students differ from that of experts?
2. What criteria do students use to decide whether two results agree? Do students consider the spread of the data, the average, or both when deciding? Students often say that results agree if they are "close". Does their judgment depend on the number of significant figures, the magnitude, relative difference, or something else?
3. Is there a particular representation that helps students correctly decide on agreement?

4.7.1 Criteria for Judging Agreement

The experts and reference sources on error analysis do not agree on the criteria used to decide if two results are consistent with each other. In fact, some references (e.g. ISO Guide, Bevington) do not even address this critical issue and leave the judgment to the reader. One simple criterion is that results are consistent when their uncertainty ranges overlap, and they are discrepant when their uncertainty ranges are not close to overlapping

(Taylor 1997). This "overlap criterion" also emerged as the most common viewpoint among the 25 experts who responded to the Measurement Uncertainty Survey and 12 new graduate Teaching Assistants in the Department of Physics and Astronomy at UNC-CH (see table below).

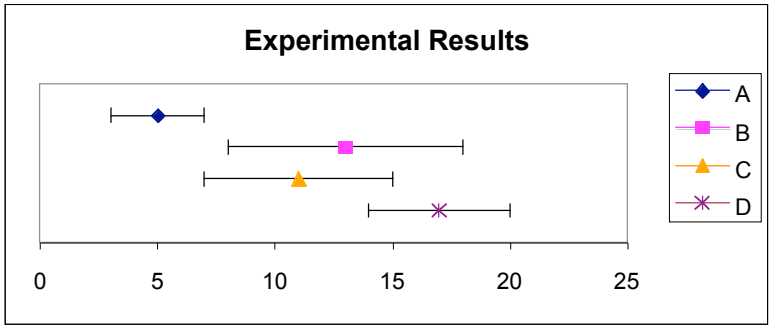
Table 4-19. Expert criteria for deciding agreement between measurements

Criteria	Expert Survey Respondents (n = 25)	UNC TAs (n = 12)
if uncertainty ranges overlap	10	8
if 1 overlap	4	
less than 2 standard errors	3	1
if difference < 3		1
use t-test (with pooled variances)	3	
not sure	2	1
other	3	1

A short survey was designed to further analyze students' criteria for agreement between measured values. This Data Comparison Survey (Appendix J) was administered to two small groups of TAs (n = 11) and students (n = 12) at UNC. Four measured values with uncertainty were presented in the survey, and respondents were asked to decide if each pair agreed with each other. These values and uncertainties were carefully selected so that the six possible combinations span the various degrees of overlap. A graphical representation of the data with error bars was also shown on the survey. Normalized Gaussian distributions that correspond to each measured value are shown here for comparison purposes (but were not shown on the survey). This Data Comparison Survey addressed two main questions:

1) When do two measurements with known uncertainties agree with each other?
2) What representation is most helpful for deciding whether results agree or disagree?

This second question was answered directly, and a clear majority of students (8/11) responded that the graphical representation with error bars was the preferred notation. It is quite interesting then that *none* of the 200+ students in this study ever drew such a graph to help them evaluate whether two values overlapped. This observation is consistent with the study by Sere, et al. (1993) where *none* of the 20 students drew a graph to compare the values and the uncertainty intervals.



A = 5 ± 2
B = 13 ± 5
C = 11 ± 4
D = 17 ± 3

Figure 4-3. Comparison of results with error bars

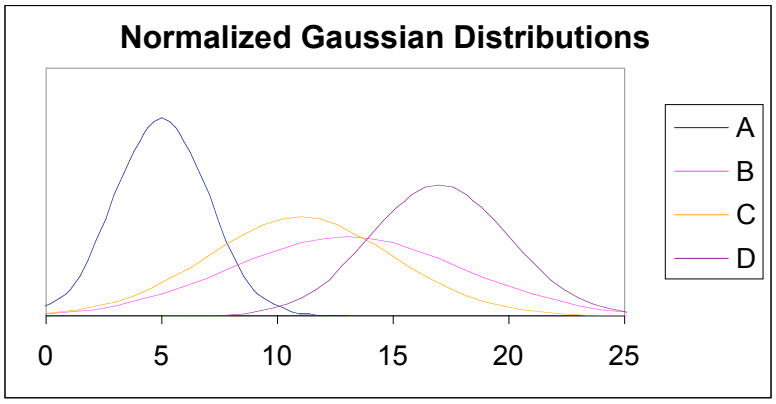


Figure 4-4. Corresponding Gaussian distributions

Note: This figure was not shown on the survey.

Students and TAs were allowed approximately ten minutes to consider their answers to each question and write them on the survey. A summary of their responses is shown in Table 4-20.

Table 4-20. Do these measurements agree?

Comparison	Degree of Overlap	UNC Students			UNC TAs	
		Y	N	?	Y	N
B = C?	Ranges overlap both midpoints	9	1	1	10	0
B = D?	One midpoint within other range	4	3	5	10	0
C = D?	Ranges overlap but not midpoints	3	6	3	9	2
A = C?	Ranges meet but no overlap	3	5	4	5	4
A = B?	Almost overlap	0	12	0	0	11
A = D?	No overlap	1	11	0	0	11

Y = yes, they agree; N = no, they do not agree; ? = not sure, more information needed

From the above table, the criteria for agreement between two results appear to depend on the degree of overlap between the uncertainty ranges. This criterion is more clearly defined in the responses from the TAs than in those of the students, who were more likely to say that an overlapping pair did not agree, but who were also more uncertain of their answers (a closer examination of the student criterion is presented in section 4.7.3). The borderline case is where the ranges just meet but do not overlap, as seen from the 5 to 4 split in opinion from the TA respondents. This borderline case is examined in the following section.

4.7.2 Overlapping Uncertainties versus t-test

While it is easy to identify uncertainty ranges shown by error bars that do or do not overlap, this criterion for agreement has several hidden complications that make it much less

clear than might be expected. An uncertainty value can represent a variety of different meanings, so an evaluator must ask a few basic questions: What confidence interval does each uncertainty represent? Do the error bars indicate some multiple of the standard deviation, the standard error, or the standard uncertainty? How many degrees of freedom are associated with each uncertainty, or what was the sample size? Are the uncertainties of the measurements being compared similar in size, or is one much bigger than the other? Is it appropriate to assume that the point estimates come from normal population distributions? Are the measurements correlated so that the uncertainties are not independent of each other? Each of these factors can affect the conclusion made from a comparison between two values and their uncertainty.

When conducting statistical hypothesis tests, two point estimates are considered *significantly different* if the test statistic indicates sufficient evidence against the null hypothesis (H_o) that the two values are equal. This evidence is given by the probability (*p*-value) that the test statistic would take a value as extreme or more extreme than the actually observed outcome (Moore 1995). If the p-value is as small or smaller than a specified significance level α, then the data are statistically significant at level α. A common significance level for general hypothesis testing is $\alpha = 0.05$ (Agresti and Finlay 1997). Some experts say that if the *p*-value is less than 0.01, there is a *highly significant difference* between the values (Taylor 1997). The *z*-test statistic is used with the standard normal distribution to compare two mean scores with known variances. An interesting question then arises: Is the *z*-test statistic with a significance level of $\alpha = 0.05$ consistent with the overlap criterion? Answering this question requires two key assumptions:

1. The measurement uncertainty represents the 68% confidence interval corresponding to $x \pm 1\sigma$.
2. The sampling distribution for x is approximately normal (not skewed).

Each of the six possible pair-wise comparisons from the Data Comparison Survey is listed in Table 4-21 and accompanied by a z-test probability for the specific values and uncertainties given, along with a range of probabilities for the degree of overlap category. The z-score is calculated from the difference between the results and the pooled uncertainty, which is then used with the standard normal distribution to find the two-tailed probability that z could be greater than the absolute value of this critical value.

$$z = \frac{x_2 - x_1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \qquad \text{and} \qquad p = 2P(Z \geq |z|)$$

The range of p-values for each degree of overlap is provided in parentheses after the specific p-value calculated for the given data (assuming that the midpoint is a mean and the uncertainty is a standard error).

Table 4-21. Probability corresponding to degrees of overlap

Comparison	Degree of Overlap	p-value
B = C?	Ranges overlap both midpoints	0.76 (0.5 to 1.0)
B = D?	One midpoint within other range	0.49 (0.32 to 0.9)
C = D?	Ranges overlap but not midpoints	0.23 (0.16 to 0.37)
A = C?	Ranges meet but no overlap	0.18 (0.16 to 0.32)
A = B?	Almost overlap	0.14 (0.1 to 0.2)
A = D?	No overlap	0.001 (0 to 0.1)

The borderline case where the uncertainty ranges barely overlap corresponds to a p-value of about 0.2 (actually 0.16 to 0.32). This means that the overlap criteria used to determine agreement will result in a Type I decision error occurring about 20% of the time, which is more frequent than the typical $\alpha = 0.05$ significance level that is used for comparing mean

values in statistical hypothesis testing. (A Type I error occurs when a researcher concludes that two values are significantly different when in fact they are not.) From the range of *p*-values, it appears that the $\alpha = 0.05$ cutoff approximately corresponds to the situation where the $\pm 1\sigma$ uncertainty ranges come close to overlapping, but do not meet.

Since there appears to be a common belief that experimental values agree if their uncertainty ranges overlap, it might be reasonable to suggest that a certain coverage factor *k* be used to expand the experimental error bars to be consistent with the widely-used $\alpha = 0.05$ significance level. As stated earlier, if the error bars represent $\pm 1\sigma$, then a Type I error will be made about 20% of the time. However, the risk of a Type I decision error could be reduced to $\alpha = 0.05$ if the error bars represent $\pm k\sigma$, such that when the error ranges barely overlap, the corresponding *z*-test probability would be 0.05. The *z*-statistic that corresponds to a probability of $p = 0.05$ for a two-tailed hypothesis test is $z = 1.96$. Therefore, to find the appropriate value for *k*, we use the condition that:

$$z = \frac{k\sigma_1 + k\sigma_2}{\sqrt{\sigma_1^{\,2} + \sigma_2^{\,2}}} = 1.96$$

The desired value of *k* then depends on the relative magnitude of each uncertainty, and the limiting cases occur when the uncertainties are equal in size or when one of the uncertainties is zero. These two extremes yield a desired range of:

$k = 1.39$ (when $\sigma_1 = \sigma_2$) to $k = 1.96$ (when $\sigma_1 = 0$ or $\sigma_2 = 0$)

Since this range is closer to $k = 2$ than $k = 1$, it seems that scientific or industrial disciplines which report uncertainties as $\pm 2\sigma$ are more consistent with accepted statistical interpretations than are disciplines like physics where uncertainties are typically quoted as

$\pm 1\sigma$ (see Table 2-1). It is interesting to note that the ISO *Guide to the Expression of Uncertainty in Measurement* does not specify a particular coverage factor that should be used for an expanded uncertainty, but mentions that values of $k = 2$ or 3 are common, since they correspond to approximately 95% and 99% confidence intervals for an assumed normal distribution.

A recent article in *The American Statistician* (Schenker and Gentleman 2001) examined this issue of evaluating the significance of differences between two point estimates by comparing the overlap between their 95% confidence intervals with the standard method of testing significance under the assumptions of consistency, asymptotic normality, and asymptotic independence of the estimates. The "standard method" rejects the null hypothesis at the 0.05 level if the 95% confidence interval for the *difference* between the point estimates does not contain 0. This difference interval is computed as follows:

$$(x_1 - x_2) \pm 1.96\sqrt{\sigma_1^2 + \sigma_2^2}$$

where x_1 and x_2 are the point estimates, and σ_1 and σ_2 are the standard errors associated with each point estimate. The "overlap method" rejects the null hypothesis at the 0.05 level if the 95% confidence intervals for each point estimate do not overlap. The nominal 95% confidence intervals for each point estimate are given by:

$$x_1 \pm 1.96\sigma_1$$
$$x_2 \pm 1.96\sigma_2$$

If these confidence intervals overlap, then there is no significant difference between the estimates.

The authors of this article conclude that the overlap method has lower statistical power than the standard method, especially when the point estimates have similarly-sized standard errors. If the null hypothesis is true according to the standard method (no significant difference), the overlap method rejects the null hypothesis less often (is more conservative; lower power). If the null hypothesis is false according to the standard method (a significant difference does exist), the overlap method fails to reject the null hypothesis more frequently (is more conservative; lower power). The overlap method approaches the standard method in the limit as one point estimate has a standard error that is much less than the other (assuming that the 95% confidence limits are employed). The authors acknowledge that the overlap method is simple and often convenient, but they conclude that the overlap method should *not* be used for formal significance testing. However, the analysis in this article only considers a 95% confidence interval for each point estimate. As discussed earlier in this section, a 68% confidence interval is most often used in physics, and this tends to have the opposite effect of having a Type I error more often than would occur with the standard method.

This article only examined the case for large sample sizes where the standard error is fairly well known. However, introductory physics labs often have small sample sizes (generally n = 1 to 10 data points), so the error (or uncertainty) in a measurement is not well known. In such cases where assumptions of normality are not met, the standard method is not valid and the overlap method is better justified since no judgement about significant differences can be made with high confidence. For example, with only 5 data points, the Student's t-statistic that corresponds with an $\alpha = 0.05$ is $t = 2.25$ when $\sigma_1 = \sigma_2$. In this case,

a confidence interval using $k = 2$ instead of $k = 1$ would yield overlap judgments that are more consist with the standard method.

In conclusion, the overlap method is more intuitive to both undergraduate and graduate students, especially if they have not studied tests of statistical significance. A graphical depiction of the overlapping confidence intervals also aides students in concluding whether their experimental results do or do not agree with a theoretical prediction within the uncertainty of their measurements. As discussed in the next section, students tend to make judgments about the quality of their data without even considering the uncertainty associated with the measurements. While the standard method is most accurate for evaluating the difference between large sample averages, the overlap method appears to be the best option for introductory physics students to use since it provides a simple and reasonably accurate way to decide if two measurements are consistent with each other.

4.7.3 Case Study for Judging Agreement

An effort was made to replicate the findings from a previous study where students were confronted with a situation where there is not clear agreement between two data sets. In the study conducted by S. Allie, et al., 121 students were asked to defend one of two positions taken in a scenario where a ball is allowed to roll down a ramp and fall onto the floor a distance d from the edge of the table:

> Two groups of students compare their results for 5 releases of a ball at $h = 400$ mm.
> Group A: 441 426 432 422 444 Average = 433 mm
> Group B: 432 444 426 433 440 Average = 435 mm
> **Group A says**: "Our result agrees with yours."
> **Group B says**: "No, your result does not agree with ours."

The following table categorizes the responses given by the students:

Table 4-22. Responses from South African students about agreement of measurements

Code	Description	Number of Students (n = 121)	Yes	No
1	It depends on how close the averages are	62 (52%)	~35%	~17%
2	It depends solely on the relative spreads of the data	4 (3%)	0	3%
3	It depends on the degree of correspondence between individual measurements in the two sets	10 (8%)	unclear	unclear
4	*It depends on both the averages and uncertainties*	*34 (28%)*	unclear	unclear
5	Not codeable	11 (9%)	unclear	unclear

According to the researchers, the most prevalent idea was to compare the average values and then decide whether the averages were "close, far, or consistent." About two thirds of this Category 1 group concluded that the two averages were consistent by suggesting that "the averages might not be the same but they are only different by 2 mm, which is a very small distance." The remaining third expressed the contrary view that "433 and 435 are totally different numbers," and several students stated that "the answers aren't exactly the same, so how can they agree with each other?" It is interesting to note that the students considered the absolute difference between the average values (2 mm) instead of the relative difference between the values (0.5%). This type of thinking is consistent with novice problem solvers (Arons 1990).

Students in Category 2 expressed statements like, "the results do not agree since the uncertainty in group A will be greater than group B." Category 3 students compared individual measurements between the sets of data and typically reasoned that "the values for the two groups match almost exactly." The most sophisticated reasoning was demonstrated by about a third of the students (Category 4) who considered both the uncertainty or spread

in conjunction with the average to come to a conclusion. However, this group had some difficulty expressing their ideas, making statements like, "if we find the uncertainties in A and B the average of A will most likely fall in the range of B(av) ± B and the same will apply to the average of B to A(av) ± A," and "with every average there should be a standard deviation and chances are both will be in the same range."

This same scenario was presented in the Data Comparison Survey to a first-semester physics laboratory course at UNC, and these students gave similar responses. Of the 11 students surveyed, 8 said that the results of Group A agreed with those of Group B. As can be seen from the written responses provided below, these students possess a wide array of vague and unclear criteria for judging agreement.

a. What do you believe? Do these results agree with each other? Please explain your answer.
b. In general, what criteria do you use to decide if two measurements agree with each other?"

> 1a. Yes, they agree with each other. Both averages and sets of data are similar leading to a conclusion that the results are at least accurate and most probably precise.
> 1b. Their accuracy or closeness to each other is the criteria I would use.
>
> 2a. Yes they agree, despite the fact that in each single release, the results can vary greatly, the averages come out to be close to one another.
> 2b. How close their averages are, how precise the data of one group is compared to the other.
>
> 3a. Yes; they have almost identical average, and individual drops of both groups are within the same range.
> 3b. Consistency
>
> 4a. I think that these results do agree with each other. Compared to the large (>400 mm) distance being measured, a difference of 2 mm is not significant enough to create a discrepancy between the results. If group B is being exact, then any difference at all, even one of 1×10^{-10} mm would create a discrepancy.
> 4b. I look at how significant the difference is in relation to the magnitude of the data being measured."

5a. Yes, neither shows big discrepancies from the mean, so they agree with each other.
5b. Precision and accuracy; the difference in the averages.

6a. I would say the results agree because 3 of 5 of their numbers match and numbers which don't are within bounds of the experiment with the exception of 422.
6b. Yes, for the same reason stated above.

7a. I believe that the results agree with each other. Their expected difference is very similar. From just eyeing the data, it appears their standard deviations would overlap.
7b. First accuracy. If the results are too widely varied, I wouldn't consider them valid.

8a. The results of both groups are similar to each other; in three cases, they had the exact same number. In another comparison between one of the two remaining sets of numbers (the ones that don't match), the difference is only 1 mm. The last comparison is off by more – 11 mm. This could be due to a human error, so in all, I think that the results are in agreement.
8b. Graphing is a more precise method, the eye can catch a difference more easily.

9a. Generally the results agree with each other due to the fact that the final answers of each group is fairly close. But they do not exactly agree with each other. Therefore, it really depends on how close you want to be. Overall, they do not agree.
9b. The final average, the closeness of each individual drop, the overall spread of the drops.

10a. I believe that the data doesn't agree because the amounts vary by too much. Of course errors will occur in both labs, but a difference of 2 mm is too much.
10b. I try to decide by how much the 2 measurements differ, in order to see if they agree.

11a. No. While the averages are nearly the same, the data is not. Group B had the one "low" data point at 426 mm and it is basically that one point that makes their average even close to group A's, who have more than one data point in that general vicinity.
11b. Not only do the average data measurements have to be nearly the same, but the patterns of the individual points must also be nearly alike.

Table 4-23. Criteria used by UNC students to judge agreement

Criteria	Frequency
(arbitrary) closeness of average results	3
same or different individual data points	3

similar spread or range of data	3
small relative difference in results	2
precision or consistency of data values	2
average and patterns must match	1
2 mm difference is too much	1

Based on the above responses, the criteria used by students to determine agreement between responses is much more vague than the overlap criterion used by experts. Students make judgments about the closeness of the agreement without considering the inherent variability of the data. These judgments are based on arbitrary standards or the student's "feel" for the size of the difference between the results. This conclusion is supported by statements made by students in interviews and in their lab reports. One student explained that the percent error between an experimental and theoretical value should be less than 10%, because that is what his high school teacher had told him (thus basing his judgment on an authority figure instead of his own empirical data). Another student used a 5% cutoff limit for an acceptable percent error, since that is what he learned from his statistics class (he had confused the $\alpha = 0.05$ level of significance with the concept of percent error). Several other students simply stated that they "felt" their experimental error was acceptable because it was "small."

Students also seem to focus their attention on the agreement of individual data points rather than the general trend of the data. All of these epistemologies are distinctly different from the expert model of thinking, which considers the difference between the results in terms of the uncertainty or spread in the results for the specific situation being investigated.

Based on student lab reports, it seems that students are often reluctant or unable to make judgments about whether their results agree or disagree with similar results. Students

also have the tendency to claim that their experimental result agrees with (or even proves!) a theory, even when such claims cannot be justified in terms of the data they collected and analyzed.

4.7.4 Best Representation for Judging Agreement

In order to examine this data comparison issue from another perspective, a combination oral/written survey was administered to the NCSU PY205 SCALE-UP (first semester calculus-based physics) class of 44 students on November 3, 1999 (which was after the students' third lab of the semester). This survey was presented via a PowerPoint presentation titled, "An Examination of Scientific Data: When are two results different?" (See Appendix I for original wording of questions). No explicit instruction had been given to students prior to this survey to judge agreement between measured values, yet they had been asked to discuss in their lab reports the results they got from experiments compared with what they expected from theoretical predictions.

Below are the student responses to each of the eight questions that were asked. Following each question is an analysis of the results and an attempt to make sense of the results in comparison with the South Africa study.

> "Suppose an experiment has been conducted to examine the effect of an independent variable on the time for an object to move along a given path."
>
> Question #1. Do these results suggest a significant difference?
>
> without treatment: $t_1 = 1.86$ s
> with treatment: $t_2 = 2.07$ s
>
> | | **Student Responses (n = 44):** |
> |---|---|
> | • YES - explain why | **19 (43%)** – "correct" response |
> | • NO - explain why | 12 (27%) |
> | • CAN'T TELL - explain why | 13 (30%) |

Since the difference between these values (0.21 s) is much greater than the precision of either value (0.01 s), there is clearly a discrepancy between the values as they are stated with no explicit uncertainty. Despite this fact, about a third of the 44 students answered that there is no significant difference between these values. Evidently these students assumed some degree of uncertainty for each value and decided that the absolute difference of 0.21 s, or the relative difference of 11%, was not large enough to consider these values to be significantly different. This type of reasoning is clearly much different from that of experts who judge agreement or disagreement between values based on the amount of uncertainty associated with the values.

Question #2. What if 2 more trials were run? Does $t_1 = t_2$?

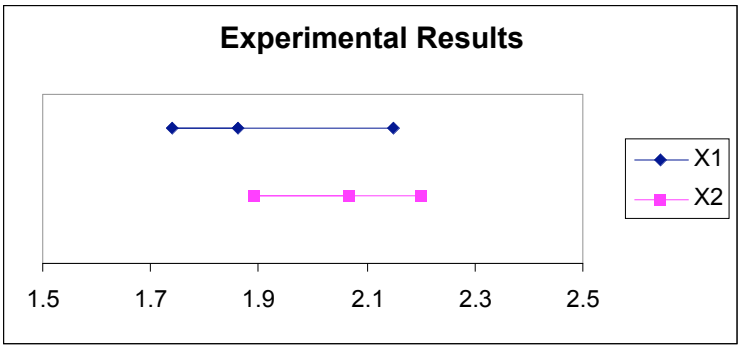
Trial #	t_1 (s) w/o treatment	t_2 (s) with treatment
1	1.86	2.07
2	1.74	1.89
3	2.15	2.20
Averages:	**1.92**	**2.05**

	Student Responses (n = 44):
• YES - explain why	10 (23%) – "correct" response
• NO - explain why	22 (50%)
• CAN'T TELL - explain why	10 (23%)
• Other	2 (4%)

Here the two data sets agree because they overlap almost entirely, although this overlap is not entirely obvious by simply glancing at the values in the data table. (The mean values also agree statistically since a t-test yields a p-value of 0.42, which is hardly sufficient evidence to reject the null hypothesis.) It would be easier to visualize this overlap if the data ranges were presented graphically, which is the purpose of the next question. It is

interesting that many students maintained their same position as they did for the first question – the values are not equivalent, despite the additional information about the variation in these numbers. This finding is somewhat consistent with the S. Africa study where 17% of the students said that the two average values they were examining (433 and 435) were not equivalent.

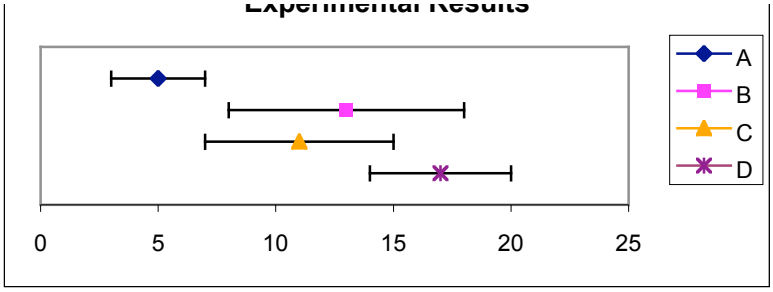
Question #3. Does $t_1 = t_2$ when plotted?



Student Responses (n = 44):

- YES - explain why — 14 (32%) – "correct" response
- NO - explain why — 26 (59%)
- CAN'T TELL - explain why — 3 (7%)
- Other — 1 (2%)

It is not clear why most of the students responded that the two values were not equivalent since the general consensus from both students and experts is that two results agree with each other when their uncertainty ranges overlap. Perhaps this question is perceived differently from Question #3 above, or maybe the students answering the questions were confused about the series of questions.



4.7.5 Conclusions about the Agreement of Measured Results

Multiple methods were used to examine student and expert thinking about the agreement between measurements. From this analysis, it appears that novices tend to think more in terms of absolute differences between results and ignore the uncertainty of the values while experts think more in terms of relative differences and also consider the uncertainty of the values when making comparisons. The "overlap criterion" is most commonly used by physics laboratory instructors to decide whether two results agree with each other. More advanced experts use the statistical *t*-test to decide agreement, but the conclusions found from applying a *t*-test do not always correspond with the overlap condition. The criteria used by students to judge agreement is often arbitrary and not as clearly defined as that of experts. Students often claim that their results agree with a theoretical prediction even when such a claim cannot be justified by the uncertainty of their experimental data.

5 Summary

5.1 Overview

The objective of this broad-based study was to examine the fundamental practices demonstrated by students related to the treatment of uncertainty associated with measurements. The research was guided by the following questions:

1. What are the common conceptions and practices demonstrated by introductory physics students regarding measurement uncertainty and error analysis?
2. How do students treat the uncertainty in measurements differently than experts?
3. Why do students believe what they do about measurement uncertainty?

5.2 Principle Findings from Students

Through this investigation, the following principle findings were discovered. These findings are part of a comprehensive list that is presented in the Appendix.

- **Arbitrary evaluation of results without considering uncertainty** - Students often make arbitrary judgments about the agreement between results and fail to consider the uncertainty estimates when making these comparisons. It appears that students do not recognize that the primary reason for determining the uncertainty in measurements is to convey the quality of the result and to make objective decisions about the agreement between results.
- **Failure to report uncertainty** - Students tend to avoid specific statements that quantify the uncertainty of a measurement, even when they are explicitly instructed to include an uncertainty estimate. This reluctance is more pronounced for directly measured values than for calculated values.

- **Failure to identify primary source of error** - Students have difficulty identifying the primary source of error in an experiment, and they generally do not analyze the effects of individual uncertainty contributions to the total uncertainty of the result. Instead, students often list a variety of possible factors that might have affected the experimental results, but these factors are rarely quantified or ranked to indicate which ones are most significant.
- **Improper use of significant figures** - Students tend to overstate precision (too many significant figures) of calculated values, and understate the precision of directly measured values.
- **Improved but limited expertise with increased exposure** - The quality of responses to measurement questions was generally aligned with the amount of training and exposure students had to the subject. While this finding is not surprising, it provides encouragement that instruction does appear to make a difference. However, even graduate teaching assistants made many of the same mistakes or omissions that were common among student responses, which suggests that these issues are not trivial to learn and apply. This conclusion is also supported by comments made by instructors who were reluctant to call themselves "experts" in the subject of error analysis.

5.3 Additional Findings

While the focus of this research was an examination of introductory physics students' understanding of measurement uncertainty, several important findings outside this scope were discovered and should be highlighted:

- In 1993, the International Standards Organization (ISO) published a new set of guidelines for expressing the uncertainty in measurements. Nearly all of the physics instructors and students surveyed in this study were unfamiliar with these recommended procedures that are now widely accepted and practiced throughout the world in industries that strive to be ISO certified. The ISO recommended practices should be incorporated into academic curricula to better prepare students for careers in science and industry.
- The notation used to express the uncertainty in measurements varies considerably among experts, which was the reason for the introduction of the ISO guidelines cited above. This inconsistency gives students confusing and conflicting instructional examples, and can result in distinctly different conclusions when comparing two values with error estimates, depending on the interpretation of the confidence level associated with each uncertainty.
- Despite the numerous possible confidence levels implied by error bars, a clear majority of both students and instructors use the "error bar overlap" criterion to decide if two results agree with each other. This finding emerged from the research (a result of the grounded theory approach) and was not expected *a priori*. The consequence of this widely used criterion is that it results in a Type I error 16% to 32% of the time if the error bars represent ± 1σ. This means that students will conclude that two results are significantly different more often with this overlap criterion than they would using a t-test with 5% significance level.
- Physics instructors reported that they learned to analyze measurement errors primarily from studying or teaching undergraduate laboratory classes. In fact, the undergraduate

laboratory experience was cited twice as often as any other source for learning error analysis. This is strong motivation to ensure that students learn proper procedures for expressing uncertainties early in their academic careers, rather than postponing introduction of this subject until advanced undergraduate labs or graduate school.

5.4 Questions for Future Research

As with all scientific research, one interesting question naturally leads to other new pathways that could be investigated. Below are several questions that were raised but not fully answered in this study.

- *Why* do students believe what they do about measurement errors?

While this question is one of the three original research questions, this study primarily examined *how* students treat errors in measurement. Explanations for some of the student difficulties have been presented here, but further examination is needed to better understand the rationale behind the student practices.

- How effective is the graphical error bar representation at getting students to use the uncertainty of their measurements to draw a valid conclusion about the agreement or difference between two values?

Although the "overlap method" is commonly employed by physics students and instructors, hardly anyone in this study used error bars to visually examine the overlap between uncertainty ranges. An Excel spreadsheet has been developed to easily allow

students to enter measurement and uncertainty values and see these error bars (Appendix x). This Data Comparison tool is now available to physics laboratory students at UNC, and the effectiveness of this tool should be evaluated.

- How frequently do students find that their experimental result does not agree with the theoretical value?

Based on the limited data obtained for this study, it appears that students often obtain experimental uncertainties that underestimated, so that a Type I error occurs more frequently than 32% of the time (as expected for an experimental value with a 68% confidence interval compared to a theoretical value with negligible uncertainty). If this perception is correct, why does it occur and should it be corrected by having students use a 95% confidence interval (or some other confidence interval) to estimate experimental uncertainties?

- How close together must two results be for students to decide they agree?

From this study, it was discovered that students often ignore the uncertainty of a measurement when evaluating a result, and they use arbitrary criteria to decide if a result is acceptable. Learning more about the students' evaluation criteria would be beneficial for developing instructional strategies to correct this common occurrence.

- If research-based curricula is developed and implemented, how effective will it be in helping students make the transition from novice to expert treatment of uncertainty?

Now that we have a basic understanding of the challenges facing students' understanding of measurement uncertainty, new curricula can be developed to address these problem areas, as has been done with other subjects through research in physics education. Implementation and evaluation of the curriculum is part of the continuing research - curriculum development - instruction cycle.

5.5 Concluding Statement

The findings from this study reveal that students have difficulties with many of the fundamental aspects related to measurements and the comparison of measured values. The most significant of these are the reluctance to specify a quantitative estimate of the uncertainty in a measured value, the inability to identify the primary source of uncertainty in an experimental result, and the failure to consider the uncertainty of a result when comparing measured values. While these are important findings, they are secondary to more fundamental problems that students have with making accurate measurements and analyzing data. Hopefully the research documented in this study will help educators improve instruction of this subject that is fundamental to all types of scientific investigations.

References

Agresti, A. and B. Finlay (1997). Statistical Methods for the Social Sciences. Upper Saddle River, NJ, Prentice Hall.

Allie, S., A. Buffler, et al. (1998). "First-year physics students' perceptions of the quality of experimental measurements." International Journal of Science Education **20**(4): 447-459.

ANSI/NCSL, Ed. (1997). ANSI/NCSL Z540-2-1997: American National Standard for Expressing Uncertainty - U.S. Guide to the Expression of Uncertainty in Measurement. Boulder, CO, National Conference of Standards Laboratories.

Arons, A. (1990). A Guide to Introductory Physics Teaching. New York, Wiley.

Baird, D. C. (1995). Experimentation: An Introduction to Measurement Theory and Experiment Design. Englewood Cliffs, N.J., Prentice Hall.

Barford, N. C. (1958). Experimental Measurements: Precision, Errors, and Truth. Reading, MA, Addison-Wesley.

Bauman, R. and C. E. Swartz (1984). "Treatment of Errors and Significant Figures in the Laboratory Manuals." The Physics Teacher **22**(4): 235-36.

Beers, Y. (1958). Introduction to the Theory of Error. Reading, MA, Addison-Wesley.

Bernstein, D. (1993). Hauling junk science out of the courtroom. Wall Street Journal.

Bevington, P. R. and D. K. Robinson (1991). Data Reduction and Error Analysis for the Physical Sciences. New York, McGraw-Hill.

Blasiak, W. (1983). "Errors and Uncertainty in Physics Measurement." Physics Education **18**(3): 290-94.

Bloom, B. (1956). Taxonomy of Educational Objectives Handbook: Cognitive Domain. New York, Longmans Green.

Box, G. E., Hunter, W. G., and Hunter J. S. (1978). Statistics for Experimenters. New York, N.Y., John Wiley.

Braddick, H. J. J. (1954). The Physics of the Experimental Method. London, Chapman and Hall.

Caballero, J. F. and D. F. Harris (1998). "There Seems To Be Uncertainty about the Use of Significant Figures in Reporting Uncertainties of Results." Journal of Chemical Education **75**(8): 996.

Chi, M. T. H., P. J. Feltovich, et al. (1981). "Categorization and representation of physics problems by experts and novices." Cognitive Science **5**: 121-152.

Chi, M. T. H., R. Glaser, et al. (1983). Expertise in Problem Solving. Advances in the Psychology of Human Intelligence. R. Sternberg. Hillsdale, New Jersey, Lawrence Earlbaum Associates. **1:** 7-75.

Clifford, A. A. (1973). Multivariate Error Analysis - A handbook of error propagation and calculation in many-parameter systems. New York, Halstead Press.

Cothern, C. R. and M. F. Cothern (1980). "Uncertainty--We Do Need It." American Biology Teacher **v42 n1**(Jan): 58-60.

Crummett, B. (1990). "Measurements of Acceleration Due to Gravity." The Physics Teacher **28, n5**(May): 291-95.

de Hoog, F. R. and C. L. Jarvis (1973). Error, Approximation, and Accuracy. St. Lucia, Queensland, University of Queensland Press.

Deacon, C. (2000). The Treatment of Numerical Experimental Results, Memorial University of Newfoundland, Department of Physics and Physical Oceanography.

Deacon, C. G. (1992). "Error Analysis in the Introductory Physics Laboratory." The Physics Teacher **30**(6): 368-70.

Deming, W. E. (1944). Statistical Adjustment of Data, Wiley.

Deming, W. E. and R. T. Birge (1934). "On the Statistical Theory of Errors." Reviews of Modern Physics **6**(July): 119-161.

Department of Physics and Astronomy, U. o. N. C. (1996). Physics 24L Lab Manual, The University of North Carolina at Chapel Hill.

Dietrich, C. F. (1991). Uncertainty, Calibration and Probability. Bristol, Adam-Hilger.

Dixon, W. J. and J. F. J. Massey (1983). Introduction to Statistical Analysis. New York, McGraw-Hill Book Company.

Doran, R. L. (1980). Basic Measurement and Evaluation of Science Instruction. Washington,
DC, National Science Teachers Association.

Edwards, M. H. (1989). “Linear Regression Revisited.” The Physics Teacher **27**(4): 280-81.

Fuller, W. A. (1987). Measurement Error Models. New York, John Wiley.

Gall, M. D., W. R. Borg, et al. (1996). Educational Research: An Introduction. White Plains, NY, Longman Publishers.

Garrison, D. H. (1975). “Random Error Experiment for Beginning Physics Laboratory.” Physics Teacher **13**(6): 356-358.

Giordano, J. L. (1997). “On the Sensitivity, Precision and Resolution in DC Wheatstone Bridges.” European Journal of Physics **18**(1): 22-27.

Good, R. H. (1996). “Wrong Rounding Rule.” The Physics Teacher **34**(3): 192.

ISO, Ed. (1993). Guide to the Expression of Uncertainty in Measurement. Switzerland, International Organization for Standardization.

ISO (1993). International Vocabulary of Basic and General Terms in Metrology (VIM). Geneva, Switzerland, International Organization for Standardization (ISO).

J. Ross Macdonald, W. J. T. (1992). “Least-squares fitting when both variables contain errors: pitfalls and possibilities.” American Journal of Physics **60, 11**(Jan): 66-73.

Jones, P. D. and T. M. L. Wigley (1990). “Global Warming Trends.” Scientific American **263**(2): 84-91.

Joram, E., K. Subrahmanyam, et al. (1998). “Measurement Estimation: Learning to Map the Route from Number to Quantity and Back.” Review of Educational Research **68**(4): 413-449.

Kurnas, S. S. (1984). “Errors with a Difference.” Clearing House **v58 n3**(Nov): 108-110.

Kuzmak, S. D. and R. Gelman (1986). “Young Children's Understanding of Random Phenomena.” Child Development **57, 3**(Jun): 559-566.

Larkin, J. H. (1981). Understanding and Problem Solving in Physics. Boulder, Co., Research in Science Education: New Questions, New Directions.

Laws, P. W. (1997). Workshop Physics Activity Guide: Core Volume with Module 1. New York, John Wiley & Sons, Inc.

Lichten, W. (1999). Data and Error Analysis. Upper Saddle River, NJ, Prentice Hall.

Lindberg, V. (1997). Uncertainties Error Propagation Graphing & Vernier Caliper: A reference manual for University Physics Laboratories, Rochester Institute of Technology.

Lubben, F. and R. Millar (1996). "Children's ideas about the reliability of experimental data." International Journal of Science Education **18**(8): 955-968.

Lyon, A. J. (1980). "Rapid Statistical Methods: Part 2a--Error Estimates." Physics Education **15**(5): 280-85.

Lyons, L. (1991). A Practical Guide to Data Analysis for Physical Science Students, Cambridge University Press.

Mandel, J. (1964). The Statistical Analysis of Experimental Data. New York, Dover.

Margenau, H. and G. M. Murphy (1947). The Mathematics of Physics and Chemistry, Van Nostrand.

McDermott, L. C. and E. F. Redish (1998). "Resource Letter on Physics Education Research." The American Journal of Physics **submitted in July 1998, not published?**

Merriam-Webster (2000). Merriam-Webster's Collegiate Dictionary. Springfield, MA, Merriam-Webster.

Meyer, S. L. (1975). Data Analysis for Scientists and Engineers, John Wiley.

Minstrell, J. (1992). Facets of Students' Knowledge and Relevant Instruction. Research in Physics Learning: Theoretical and Empirical Studies. F. G. R. Duit, & H. Niedderer. Kiel, Germany, IPN**:** 110-128.

Moore, D. S. (1995). The Basic Practice of Statistics. New York, W. H. Freeman and Company.

Mueller, J. W. (1984). Precision Measurement and Fundamental Constants II. National Bureau of Standards Special Publication 617. B. N. a. P. Taylor, W. D. Washington, D. C., US GPO**:** 375-381.

Mulliss, C. L. and W. Lee (1998). "On the Standard Rounding Rule for Multiplication and Division." Chinese Journal of Physics **36**(3): 479-487.

Natrella, M. G. (1963). Experimental Statistics. Washington, U.S. Dept. of Commerce, National Bureau of Standards, U.S. Govt. Print. Office.

NPR (1997). National Public Radio - All Things Considered news program.

NRC (1996). US TIMSS Summary of Findings, U. S. National Research Center.

Osborne, J. (1996). “Untying the Gordian knot: diminishing the role of practical work.” Physics Education **31**(5): 271-278.

Paulos, J. A. (1988). Innumeracy: Mathematical Illiteracy and Its Consequences. New York, Hill and Wang.

Pfunot, H. and R. Duit (1994). Bibliography: Students' Alternative Frameworks and Science Education. Kiel, IPN.

Phillips, M. D. (1972). “A Simple Approach to Experimental Errors.” Physics Education **7**(6): 33-388. Physics, H. P. Measurement and Precision, Experimental Version. Cambridge MA., Harvard Univ.

Poultney, S. K. (1971). “Measurement and Its Reliability: An Introductory Laboratory Experiment.” American Journal of Physics **39, 2**(Feb): 176-182.

Pugh, E. M. and G. H. Winslow (1966). The Analysis of Physical Measurements, Addison-Wesley.

Reif, F. and M. S. John (1979). “Teaching Physicists' Thinking Skills in the Laboratory.” Am. J. Phys. **47**(11): 950-957.

Roberts, D. (1983). “Errors, Discrepancies, and the Nature of Physics: An approach to physics labs.” The Physics Teacher **21**(3): 155-60.

Schenker, N. and J. F. Gentleman (2001). “On Judgind the Significance of Differences by Examining the Overlap Between Confidence Intervals.” The American Statistician **55**(3): 182-186.

Schigolev, B. M. (1965). Mathematical Analysis of Observations. London, London Iliffe Books.

Sere, M. G., R. Journeaux, et al. (1993). “Learning the statistical analysis of measurement error.” International Journal of Science Education **15**(4): 427-438.

Serway, R. and R. Beichner (2000). Physics for Scientists and Engineers. Fort Worth, Saunders College Publishing.

Soh, J., B. Fairbrother, et al. (1998). Students' Skills and Conceptions on Measuring. Seoul, Korea, Soul National University.

Spencer, C. D. and P. F. Seligmann (1992). “An Independent Freshman Laboratory.” The Physics Teacher **30,**(5): 310-14.

Squires, G. L. (1985). Practical Physics. Cambridge, Cambridge University Press.

Stanford, B. E. (1999). How Wrong Can I Be? (and what does "plus or minus" really mean?). Raleigh, NC, NC State University.

Strauss, A. and J. Corbin (1990). Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Newbury Park, CA, SAGE Publications.

Swartz, C. (1995). "Editorial: Wintry Thoughts." The Physics Teacher **33**(4): 202.

Swartz, C. (1998). "Editorial: Practically Perfect in Every Way." The Physics Teacher **36**(3): 134.

Swartz, C. (1999). "Editorial: The Error of Our Ways." The Physics Teacher **37**(10): 388.

Swartz, C. and T. Miner (1997). Teaching Introductory Physics: A Sourcebook. Woodbury, NY, AIP Press.

Swartz, C. E. (1993). Used Math for the first two years of college science. College Park, MD, American Association of Physics Teachers.

Sydenham, P. H. (1983). Handbook of Measurement Science. Chichester, John Wiley & Sons.

Tawney, D. A. (1972). "The Design of Experiments and the Estimation of Experimental Errors: A Necessary Preparation for Project Work." Physics Education **7**(6): 377-382.

Taylor, B. (1999). Physics Laboratory, National Institute of Standards and Technology.

Taylor, B. N. and C. E. Kuyatt (1994). NIST Technical Note 1297: Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. Washington, U.S. Government Printing Office.

Taylor, J. R. (1997). An Introduction to Error Analysis. Sausalito, University Science Books.

Thompson, W. (1992). "Physics, Error Analysis, and Statistics." American Journal of Physics **60, 11**(Nov): 969.

Weast, R. C. (1988). CRC Handbook of Chemistry and Physics, CRC Press.

Wilson, E. B. (1952). An Introduction to Scientific Research. New York, McGraw Hill.

Worthing, A. G. and J. Geffner (1950). Treatment of Experimental Data. New York, Wiley.

Yeatts, F. R. (1979). "Measurement Oriented Basic Physics Laboratory." American Journal

of Physics **47, n1**(Jan): 46-49.

Young, H. D. (1962). Statistical Treatment of Experimental Data, McGraw-Hill.