



A Pilot Study Teaching Metrology in an Introductory Statistics Course

[Emily Casleton](#)

Iowa State University

[Amy Beyler](#)

UnitedHealth Group

[Ulrike Genschel](#)

Iowa State University

[Alyson Wilson](#)

North Carolina State University

Journal of Statistics Education Volume 22, Number 3 (2014),
www.amstat.org/publications/jse/v22n3/casleton.pdf

Copyright © 2014 by Emily Casleton, Amy Beyler, Ulrike Genschel, and Alyson Wilson, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Variability; Statistics curriculum; Undergraduate statistics; Deep understanding; Data collection.

Abstract

Undergraduate students who have just completed an introductory statistics course often lack deep understanding of variability and enthusiasm for the field of statistics. This paper argues that by introducing the commonly underemphasized concept of measurement error, students will have a better chance of attaining both. We further present lecture materials and activities that introduce metrology, the science of measurement, which were developed and tested in a pilot study at Iowa State University. These materials explain how to characterize sources of variability in a dataset, in a way that is natural and accessible because the sources of variability are observable.

Everyday examples of measurements, such as the amount of gasoline pumped into a car, are presented, and the consequences of variability within those measurements are discussed. To gauge the success of the material, students' initial and subsequent understanding of variability and their attitude toward the usefulness of statistics were analyzed in a comparative study. Questions from the CAOS and ARTIST assessments that pertain to using variability to make

comparisons, understanding the standard deviation, and using graphical representations of variability were included in the assessment. The results of the comparative study indicate that most students who were exposed to the material improved their understanding of variability and had a greater appreciation of the value of statistics.

1. Introduction

An important task in introductory statistics courses is to help students develop an understanding of the concept of variability in data and to illustrate how variation is at the core of statistical reasoning and critical thinking. As instructors, we emphasize that variability exists in virtually all data; we present methods to graphically and numerically describe and summarize variation; we then proceed to demonstrate how variation is essential for statistical inference. Nevertheless, instructors often find it difficult to successfully relate the concept of variability to students ([Reading 2004](#)). Extensive research has shown that gaining a *deep* ([Garfield and Ben-Zvi 2005](#)) and *robust* ([Peters 2011](#)) *understanding* of variability is more complex and difficult than previously perceived ([Wild and Pfannkuch 1999](#); [Reading 2004](#); [Reading and Shaughnessy 2004](#)). Instructors are often confronted with pre-existing misconceptions about variability that students bring to the classroom ([Hammerman and Rubin 2004](#); [delMas and Liu 2005](#); [Garfield and Ben-Zvi 2005](#); [Makar and Confrey 2005](#)); we discuss specific examples in Section 2.2. One purpose of this study is to assess the usefulness of introducing metrology in an introductory-level course in order to improve students' understanding of variability and to clarify existing misconceptions about the topic.

Further common challenges in introductory statistics courses are often a lack of interest by students in statistics itself ([Lee 2007](#); [Keeley, Zayac, and Correia 2008](#)) and negative preconceptions and attitudes toward statistics ([Gal and Ginsburg 1994](#)). This is of concern, as behavioral research suggests that the less students relate to a subject, the less they will learn (cf. [Anderson, Shirey, Wilson, and Fielding 1987](#); [Schiefele 1991](#); [Weber, Fornash, Corrigan, and Neupauer 2003](#)). Although students are frequently exposed to statistics in everyday life, they often do not recognize it. As a consequence, they do not see the relevance of statistics to their personal life, nor their field of study or profession. Failing to recognize the practical importance of statistics can lead students to miss the usefulness of the course material ([Snee 1993](#)), which ultimately results in a lack of motivation and interest ([Schiefele 1991](#)).

We chose *metrology*, the science of measurement, as a possible topic for statistics courses because (1) measurement issues arise not only in our personal lives, but virtually in all sciences and many other areas, such as manufacturing, quality control, health care, public safety, environmental assessment, and agriculture, and (2) the applications can be used to motivate a broad range of both introductory and advanced statistical concepts. We present lectures and group activities suitable to introduce students to the concept of metrology in an introductory-level statistics course for majors and non-major students. We illustrate measurement error through relatable examples to serve as a vehicle for demonstrating variability and to help improve students' perception of the field of statistics.

The remainder of this paper is organized as follows: Some background and misconceptions about metrology that have motivated this work are presented in Section 2. The learning objectives are

stated in Section 3. Section 4 describes some of the features of the lab and lecture material. Section 5 describes the methodology used to test the developed material. Results of the assessments are presented in Section 6. Discussion and conclusions are included in Section 7.

2. Background and Motivation

2.1 Metrology and the current undergraduate curriculum

Metrology, the science of measurement, is an underemphasized topic in most statistics courses ([Vardeman et al. 2010](#)). It is fair to say that the majority of students are unfamiliar with the idea of measurement error. As an example, suppose that we randomly select ten one-pound pre-packaged tubes of ground beef from our local grocery store. We weigh each tube. Depending on the precision of our scale, it is likely that we will obtain ten slightly different measurements, all of which differ from the one-pound nominal weight. If we pick one particular tube and weigh it again, perhaps on a different scale, we will almost always observe a different value. The variability in the measurements can come from different sources. For example, the scale used during the filling process of the tubes is different from the one we used. Our scale is likely to have a different calibration, and it may be less precise than the one used by the butcher. Many students would measure the weight of a tube only once and take that weight to be exact. If students take repeated measurements of the same tube and end up with different measured values, they might assume that one or more of the measurements are incorrect. One of the most common misconceptions ([Allie, Buffler, Campbell, and Lubben 1998](#); [Deardorff 2001](#)) is that the correct weight of an item is the measurement that was obtained repeatedly through multiple measurements. Results of data analyses, of course, are only as accurate as the data used in them. Because data analysis is at the center of most introductory level statistics courses, it seems natural to address the existence of measurement error and how it plays into data analysis there.

[Garfield and Gal \(1999\)](#) assert that the notion of measurement error is one of the “big ideas” for understanding statistical investigation, which they list as a widely accepted learning goal for statistics students. Nevertheless, the topic has not received much attention, which might simply be a result of an already dense curriculum in statistical training at both the K-12 and college levels. Other disciplines, such as physics and chemistry, have already recognized the importance of this topic. In these sciences, published, introductory-level learning materials are available to assist students in learning about measurement error and measurement variability (see, for example, [Deardorff \(2001\)](#), [Pillay, Buffler, Lubben, and Allie \(2008\)](#), or [Buffler, Lubben, Allie, and Campbell \(2009\)](#) in introductory physics or [Zipp \(1992\)](#), [Kimbrough and Meglen \(1994\)](#), or [Prilliman \(2012\)](#) in chemistry).

Measurement variability is direct evidence of the statistical concept of variability. If students become aware of the existence of measurement variability, they satisfy one of the fundamental principles of understanding variation: the reality of variation in all data (cf., [Moore 1990](#); [Wild and Pfannkuch 1999](#); [Garfield and Ben-Zvi 2005](#); [Reid and Reading 2008](#); or [Peters 2011](#)). [Wild and Pfannkuch \(1999\)](#) include sources of variation such as measurement devices and operators (“Measurers,” p. 235) as part of the ever-present variability in data. [Peters \(2011, pp. 54\)](#) lists the “anticipation of possible sources of variability (such as measurement variability)” as one of four items necessary to reason about variation in the context of observational and experimental

design. [Reid and Reading \(2008\)](#), propose a four level “Consideration of Variation Hierarchy.” The third level, “*Developing* consideration of variation,” aligns with main metrological ideas such as recognizing and describing within-group variation (e.g., several measurements obtained using the same measurement device and operator) and between-group variation (e.g., measurements obtained using more than one operator or device). At a more advanced level of statistics, metrology can also be used to teach the connection between within-group and between-group variation for the purpose of making statistical inference, thus satisfying what Reid and Reading define as the highest level of the hierarchy, namely “*Strong* consideration of variation.”

[Garfield and Ben-Zvi \(2005\)](#) discuss and organize the main ideas and propositions put forward by several authors in a series of articles on reasoning about variation (see [Garfield and Ben-Zvi 2005](#), p.93, for references therein and for more information). They provide a list of several components necessary for achieving a “deep understanding” of variability. We argue again that most of these components can be addressed using metrology examples. Specifically, *developing intuitive ideas of variability*, is considered by illustrating that variability occurs naturally in the measurement process and may have different causes, e.g. multiple operators, multiple measurement devices, or measurements taken on multiple days. *Describing and representing variability* can be demonstrated and practiced with students through studying side-by-side boxplots or dotplots, for example. Side-by-side boxplots in the context of metrology are valuable for displaying a set of measurements separately by operators, machinery or other factors known, or having the potential, to affect a measurement. Side-by-side boxplots also serve as a starting point for *using variability to make comparisons*. Metrology also allows instructors to help students *consider variability as part of statistical thinking* by introducing examples that tie data analysis to daily activities. Additionally, metrology further serves as an anchor to revisit variability at a different point in the course and from a different perspective as suggested by [Garfield and Ben-Zvi \(2005\)](#). Because of this interconnection, we hypothesize that metrology can aid students to gain a deeper and more profound understanding of the concept of variability.

2.2. Students’ perceptions and understanding of the measurement process

Before data can be analyzed, they must be collected. This is an important process that is often hidden from students (e.g., to save class time or simply for convenience). This may explain why students seldom question the source, quality, or integrity of the data presented to them; students frequently accept data points as “true,” including extreme observations or data points that are physically impossible ([Allie et al. 1998](#); [Deardorff 2001](#); [Pillay et al. 2008](#)). [Séré, Journeaux, and Larcher \(1993\)](#) conducted an extensive study exploring students’ understanding of obtaining measurements in the context of optics and electricity. Their findings include that students often do not see a need for repeated measurements and when asked to take more than one measurement, students used subsequent measurements to confirm earlier measurements. These same students generally did not distinguish between random and systematic error in data and frequently did not fully understand the objective of constructing confidence intervals.

Collecting measurements, reporting and interpreting the observed values, and understanding the uncertainty resulting from measurement error is an important and recognized concept in the lab sciences, e.g. chemistry and physics, and has received considerable attention in these fields. This allows us to present known and widespread misconceptions that students may have regarding

measurements and measurement error before proceeding with the main learning objectives associated with metrology. Perhaps the most in-depth study of students' misconceptions about obtaining measurements in the context of experimental data was conducted by [Lubben and Millar \(1996\)](#) in the United Kingdom, surveying over 1000 students of ages 11, 14 and 16. The authors developed a "model of progression of ideas concerning experimental data." This model has widely been confirmed (also when applied to college level students) in later studies such as [Allie et al. \(1998\)](#), [Buffle, Allie, and Lubben \(2001\)](#), [Deardorff \(2001\)](#), or [Pillay et al. \(2008\)](#).

Examples of misconceptions and misunderstandings of measurement uncertainty in students ranging from middle school to college include (cf., [Lubben and Millar 1996](#)):

- "Measure once and this is the right value"
- "Unless you get a value different from what you expect, a measurement is correct"
- "Make a few trial measurements for practice, then take the measurement you want" (also identified as "perfecters" in [Allie et al. \(1998\)](#))
- "Repeat measurements till you get a recurring value. This is the correct measurement" (also identified as "confirmers" in [Allie et al. \(1998\)](#))
- "You need to take a mean of different measurements. Slightly vary the conditions to avoid getting the same results"
- "Take a mean of several measurements to take care of variation due to imprecise measuring. Quality of the results can be judged only by authority source (sic)"
- "Take a mean of several measurements. The spread of all the measurements indicates the quality of the result"
- "The consistency of the set of measurements can be judged and anomalous measurements need to be rejected before taking a mean"

Similar to [Lubben and Millar \(1996\)](#), [Allie et al. \(1998\)](#) identified a so-called group of "mean-reasoners," some of whom (incorrectly) repeat measurements for the purpose of improving measurement accuracy, while other students in this group take repeated measurements to either establish a mean (the majority of students) or establish a spread. Perhaps most interesting is, however, that a significantly lower number of students in this study saw the need to repeat measurements when measuring distances compared to when taking measurements of time.

The course materials that we discuss in Section 4 directly address some of these common misconceptions. We next define the specific learning objectives associated with the metrology lecture material.

3. Learning Objectives

In the following, we distinguish between a broader set of learning objectives that focus on improving students' perception, interest, and understanding of statistics as well as variability (denoted "B"), and a set of learning objectives that focus on metrology-specific issues (denoted "M"). Three broad learning objectives, ordered from a lower to higher level of understanding emerge:

- B1. To improve students' perception of the use and value of statistics in their personal and professional lives.

B2. To foster a deeper understanding of variability in data.

B3. To introduce the idea of measurement error in data.

These learning objectives are primarily intended for students who take a first statistics course that frequently serves to fulfill a quantitative reasoning requirement in many colleges and universities. The majority of these students tend to have little mathematical background and a diverse set of academic interests.

We distinguish between the following metrology-specific learning objectives:

M1. To recognize and account for measurement error even when reporting a single measurement.

M2. To examine the role of variability in the measurement process and understand how this variability influences how measurements are taken.

M3. To recognize that measurements are approximations to the true value of interest due to for example finite precision and the potential for bias in the results obtained from any measurement device.

M4. To identify features and patterns in a graphical display of data that illustrates the contributions to variability from different sources of measurement error.

The metrology-specific objectives address the experimentation-focused misconceptions discussed in Section 2.

In addition to fostering the above-mentioned learning objectives, the developed material also addresses a number of additional outcomes (denoted “A”) such as:

A1. Apply metrology definitions such as variability, measurand, measurement, device, operator, measurement error, and reference standard to specific examples.

A2. Recognize that measurement error comes from multiple sources and generate a list of common sources of variability, such as devices or operators.

A3. List ways to mitigate measurement error.

A4. Explain why identifying and controlling sources of measurement error is an important part of conducting an experiment.

4. A Set of Course Materials on Metrology

The classroom materials developed to introduce metrology include a set of lecture notes and two interactive group activities suitable for lab sessions. The lecture notes provide an overview and introduction to the concept of metrology at a level that is appropriate for any introductory statistics course at the undergraduate level. The material requires no additional prerequisites for students, which adds flexibility to the timing at which the material can be presented throughout a course. For example, the material can be presented early in the semester, following the introduction to data and principles of data collection as well as the definition of variability, but before students see a formula for standard deviation. For that timing, the metrology material can reinforce the concept of variability by focusing on its physical sources.

Alternatively, the topics can be introduced once students have encountered the terminology population, parameter, sample, statistics, measures of central tendency and spread, graphical displays, and the basic idea of a completely randomized experimental design. This was the

timing used in the pilot study, which was beneficial, as we were able to reinforce the above-mentioned concepts with analogous ideas in the metrology material.

Finally, the metrology material can be introduced once students are familiar with confidence intervals and statistical significance. Knowledge of these concepts will aid when students are taught to report single measurements as an interval and to identify significant uncertainty arising from systematic experimental sources.

In order to relate to a potentially diverse audience, we attempted to choose broad applications when explaining concepts. To better relate to students and illustrate usefulness we recommend, however, adjusting examples or data collection settings to students' specific academic backgrounds and interests whenever feasible.

All of the material was developed following several of the GAISE recommendations ([ASA 2005](#)) such as using real data and emphasizing statistical thinking and concepts rather than procedures. The group activities and assessment questions focus not only on the content knowledge of metrology, but also on more conceptual ideas, with an emphasis on variability. In the discussion below, we detail the current lecture material and group activities. Note that these materials have been updated since their original implementation in the pilot study based on student, peer, and reviewer feedback. While these changes helped clarify content and concepts, they did not alter the overall objectives of the lessons and activities. The updated materials are provided in the supplemental material linked here:

[Lecture-Part1](#)

[Lecture-Part2](#)

[LessonPlan-Activity1-Single Measurement](#)

[LessonPlan-Activity2-More than 1 Measurement](#)

[GroupActivity1](#)

[GroupActivity1-pictures](#)

[GroupActivity2](#)

4.1 Lecture 1: Measurement Error

The lecture material is divided into two parts to be introduced in two separate class sessions: the first set of lecture notes focuses on *acknowledging the existence of measurement error* as well as *reporting measurement error*. We recommend following the lecture with the accompanying activity that allows students to work in groups.

Lecture 1 begins by introducing students to the term *metrology*, which is defined as the science of measurement. The importance of this topic is emphasized by the existence of an entire field of science devoted to it. We present measurement error as an integral element of metrology and discuss the metrology-specific learning outcomes in terms of measurement error: what is measurement error, when and why does measurement error occur, why is measurement error important, and how does measurement error affect the experimentation process and the reporting of measurements? These outcomes are revisited at the end of the second lecture when students are asked the same set of questions in relation to a capstone example.

The lecture proceeds with a motivating example demonstrating the importance of considering sources of measurement error and the dangers of making a decision based solely on a single measurement without quantification of uncertainty. In this example, students are placed in the position of a patient at a doctor's visit with a single high blood pressure reading. This one measurement suggests that the student might have stage 1 hypertension, potentially requiring medication. Because such medication typically is not only a lifelong commitment with possible side effects but also has economic consequences, physicians generally base such decisions on more than one measurement from the patient. Since blood pressure is affected by many factors, such as time of day, the doctor might ask the student to return five times during the following week to have additional blood pressure measurements taken. At each visit, the blood pressure is measured using three different sphygmomanometers, i.e., devices. Of the resulting 15 measurements used in this example, ultimately only two are at or above the designated cutoff for stage 1 hypertension. Students are then asked if the medication should be prescribed if the doctor was only aware of a subset of the measurements, such as those recorded on Thursday, those taken with device #3, or only the single highest recording. The purpose of this example is heuristic; it encourages students to consider a scenario in which a single measurement, or a single value, does not take into account confounding effects that may impact the measurement and ultimately a decision. We acknowledge that this example may present challenges; for instance, identification of the measurand, a concept which is presented later in the metrology materials. However, in practice, this example was well received in the classroom when first piloted and served as a motivating example for the subsequent material.

The lecture material continues by defining the terminology necessary to develop a conceptual understanding of metrology. The term, *variability*, is defined as the degree to which a set of measured values fluctuates under fixed measurement conditions and *variance* is defined as one measure to quantify variability. It is noted that variability cannot be assessed with a single data value, foreshadowing again the need and benefit of repeated measurements. Although students have likely been introduced to the definition of variability before, reiteration is important and links immediately to two of the primary learning goals, B2 and M2. The existence of variability is demonstrated through various examples and discussion stimulating questions. Two of the examples ask students to evaluate and compare the amount of variability resulting from two different scenarios, while two other examples illustrate a complete lack of variability. None of the examples require students to quantify variability, i.e. to calculate a standard deviation or range, but rather approach these tasks conceptually.

Next, the terms *measurand* and a *measurement* are defined and discussed. The relation between a measurand – a quantity for which one hopes to establish a true or correct value – and measurement – the experimental assignment of a potential value to a measurand – can be seen as analogous to the relation between parameter and statistic. This analogy is particularly beneficial to stress if students already have seen the terms parameter and statistic. *Device* and *operator* are defined separately as the equipment and procedures used to produce the measurement and the person reading the measurement from the device, respectively. These terms are necessary as they often contribute to variability within repeated measurements, which will be discussed at length in the second lecture.

We next define *measurement error for a single measurement* as the amount by which each observed measurement differs from the true, but unknown value. It should be noted that both lectures consider measurement error only for continuous, quantitative variables. The definition of measurement error automatically dictates that measurements should be interpreted as approximations to the true value of the measurand. Because the target audience for the material are students in an introductory statistics course which typically does not have a mathematics prerequisite, we made a conscious decision to not define measurement error through formula notation, i.e., the difference between the measurement and the true value of the measurand as would be appropriate at a more technical level.

The last term defined is *reference standard*, which is a realization of a specified quantity used as a measurement unit. Typical reference standards are multiples of meaningful measurement units, such as a kilogram, an inch, or a second. Students are informed that the organizing body in the United States that develops and maintains national measurement standards is the National Institute of Standards and Technology (NIST), and that there are analogous national and international organizations that work to harmonize measurement standards, especially those used to realize the measurement units that make up the International System of Units (e.g. the kilogram, meter, etc.), across countries.

A fundamental idea and emphasized discussion point in this first lecture is that measurement error always exists as an unavoidable consequence of the finite precision and potential bias present in all measuring devices. This idea is stressed to specifically counter the general misconceptions introduced in Section 2.2. Students are encouraged to understand measurement error not as an error in the typical colloquial sense but rather that measurement error is inherent to every measurement process. This discussion can further be extended to the relationship between the cost of a measurement device and the measurement quality it will be able to provide and the need to consider different degrees of accuracy or precision depending on the application.

The first lecture concludes with a discussion on how to scientifically report the result of a single measurement while at the same time accounting for measurement error. All measurements should be reported as an interval of plausible values for the measurand to account for the inherent measurement error. If the device used to make the measurement has a stated calibration or rating this value should be included when determining the interval to report. If no calibration is specified for the device, it is common practice to assume that the last digit reported is imprecise, i.e., the measurement is reported as plus/minus one in the final digit. This practice accounts for device resolution and operator precision and is borrowed from what is frequently taught in chemistry education ([Pacer 2000](#); [Terezakis 2010](#)). In addition, reporting measurements as an interval of plausible values reiterates or alludes to the idea behind confidence intervals.

4.2 Group Activity 1: Measurement Error

A group activity was developed to reinforce the main concepts presented in the first lecture. The specific learning objectives of the activity are for students to (1) recognize the omnipresence of measurement error, (2) distinguish between a measurand and a measurement, (3) report measurement results as an interval to account for measurement error, and (4) recognize the importance of precision relative to the measurand. The activity asks students to measure the

weight of two distinct substances and requires students to think about how the amount of tolerable measurement error naturally differs in the two applications.

First, students are placed into the role of a quality control worker at a Betty Crocker factory. It is the students' task to measure and check the weight, in ounces, of a random sample of manufactured boxes of a brownie mix. (Note that a similar task was first suggested by the Physics Education Group, University of Cape Town and the Science Education Group, University of York, website: <http://www.phy.uct.ac.za/people/buffler/>, p.7). Under ideal circumstances, the activity requires one or more brownie mixes, a balance and a set of different calibrated masses. The weight of each box is determined by placing calibrated masses successively on the balance and students are asked to report the weight of the mix. Although using a balance to weigh a substance may seem old-fashioned, reporting measurements as an interval is a natural result of this method. Further, the balance example easily demonstrates the finite precision of measurements, which here is determined by the smallest calibrated mass. Once the smallest mass available has been used, a more precise measurement cannot be obtained. A precocious student may recognize that we are not explicitly accounting for the inaccuracy of the weight of the calibrated masses. Although this is true, Lecture 1 does not discuss a method to incorporate multiple sources of error in reporting a single measurement. Thus, the calibration of the masses is not considered in this set of introductory materials. Because the required resources for this activity may not be feasible, we created a paper version of this activity that provides a set of illustrations of a balance with calibrated masses allowing students to complete this task.

For the second activity, students work in a pharmaceutical company. This time students are tasked with weighing and verifying the amount of acetylsalicylic acid, an active ingredient in aspirin tablets, for a test batch to be manufactured. Ideally, this activity requires a digital scale with three (or more) settings of precision and a substance that resembles acetylsalicylic acid, such as fine crystal sugar. Initially, students obtain three measurements from the digital scale, one measurement for each level of precision beginning with the level of least precision. Each new measurement then allows reiterating the important distinction between the measurement and the true value of the measurand, namely that an observed measurement is *only* an approximation to a true, but unknown, value. With each measurement students are also asked to report the single measurement while accounting for measurement error. Although a digital scale is more technologically advanced than the balance, students will see that there still is a limit to the precision that comes with this measurement device. Again, a paper illustration of this activity exists if the necessary resources are not readily available.

4.3 Lecture 2: Measurement Error with Multiple Observations

The second lecture addresses how to *interpret information from a set of measurements* and how *measurement error affects experimentation*. The lecture begins by presenting measurement scenarios students are likely to have experienced previously or can relate to: pumping a specified amount of gas at a gas station, buying bananas at a grocery store, and using a carbon monoxide monitor at home. For each scenario, students are asked to consider and discuss the consequences of neglecting the inherent measurement error. We continue with a more detailed example, which we reference throughout the first part of the lecture. To motivate this example, students are given the American Academy of Orthopaedic Surgeons' (AAOS) recommendations on backpack safety

for children and teenagers. The AAOS recommends, for example that “Kids should carry no more than 15 to 20% of their body weight.” To explore if this is a realistic assumption on college campuses students are asked to estimate the average weight of the textbooks required for their classes. The process of obtaining a sampling frame, using random selection, weighing the selected books, and calculating an average is discussed in detail in the context of this example. Building on the inaccuracy of each individual book weight, we can now illustrate to students how measurement variability influences what can be said about the average weight of the sampled textbooks.

Rather than viewing a set of measurements as a set of approximate values, students are encouraged to view measurements as capable of providing additional information. Not only do measurements inform us about the true value of the measurand, but also about the measurement process. In the context of the illustrative example, the average weight of the sampled textbooks can be interpreted as the true average weight of the textbooks plus the average amount of the scale’s inaccuracy. However, the average is only one number and does not provide any information about how to separate these two pieces of information. The discussion then turns to the variance and separating the amount of variation that can be attributed to the heterogeneity of the book weights versus the uncertainty that stems from the measurement process. This cannot be accomplished by considering the reported sample average itself but requires examining the variability found in the dataset in more detail.

The textbook data consist of 50 measurements from 50 different sampled textbooks (i.e., there are no repeated measurements at the individual book level), thus the existing data cannot be used to separate the variability into the components that are either due to the scale or the books alone. To gather such information requires a particular experiment and data collection, which students are asked to consider and design. Specifically, students are asked how to obtain a dataset such that if variability in the values was present students could be reasonably sure the variability was due to the scale. This leads into a discussion of how experimental design and data collection are affected by measurement error. A first step in accounting for measurement error in an experiment is to identify factors that potentially contribute to the overall measurement variability. There are three types of factors that typically can be identified: those which can be controlled and for which repeated measurements can be obtained, those which can be controlled but for which no repeated measurements are possible, and those which cannot be controlled. Three commonly identified factors are the measuring device, the operator, and the surrounding physical circumstances. Examples for each are discussed in detail. The previous discussion enables us now to present the overall amount of variability in the data as the sum of the partitioned variances resulting from the various sources of measurement error. This presentation is at a very conceptual level without showing students statistical formulas and calculations, which we deemed unhelpful at this level.

The lecture notes continue with the following capstone example. Note that we used JMP ([JMP 2012](#)) to analyze the data and focused primarily on the results from the corresponding variance component analysis. For the capstone example we implemented a full factorial experiment to determine the width of cereal boxes by the Healthy Cereal Company. The experiment utilizes three different operators, two different measurement devices, and 10 distinct cereal boxes. Each

cereal box is measured three times by each operator and device combination, resulting in 180 total measurements.

In the lecture notes students are asked to identify the three factors that likely contribute to variability in these data and for which the associated amount of variability is estimable: the operator, the device, and the part-to-part variation. For each factor, the lecture notes detail which subset of measurements are necessary to quantify the associated variance. Variability from the device is separated out into *within* device variability, i.e. variability among measurements obtained from the same device, and between device variability, i.e., variability among measurements that were made with a different device. If Analysis of Variance (ANOVA) is a topic covered in the course, this example and its discussion can be used to reinforce or even introduce related statistical concepts such as the set-up of the ANOVA table or the definition of the F-statistic.

The lecture notes remind the students that some amount of variability remains unexplainable, owing to those factors not considered or those for which repeated measurements were not obtained. Because focus was placed on the conceptual idea of measurement error with multiple observations and the idea of different sources of variability, we finish the presentation of this example with a pie chart showing the decomposition of the total amount of variability into its estimable components, called variance components. More emphasis can be placed on the calculations necessary to obtain the estimated variance components for students at a more advanced level. In conclusion students are asked questions related to the pie chart, such as, should the company use only one operator, and if only one thing could be improved in the entire measurement process, what should this be?

The lecture concludes by restating the learning objectives from the beginning of the metrology material within the context of the capstone example. For example, students are asked if there would still be measurement error if the operators had used more sophisticated measuring devices and how measurement error affected the design of the cereal box experiment. We discuss that failure to accurately account for measurement error may lead to a decrease in knowledge about the true value of the measurand. An example of a poorly designed experiment that does not allow any of the potential sources of measurement error to be estimated is presented. As a final task, students are encouraged to discuss how the experimental design could be improved to learn more about the measurement errors and the true value of the measurand.

4.4 Group Activity 2: Measurement Error with Multiple Observations

We developed a group activity to reiterate the main concepts presented in the second lecture. The specific learning objectives of the activity are for students to (1) understand that measurement variability affects experimental design, (2) identify potential factors that contribute to the overall measurement variability, (3) understand that the overall variability can be partitioned into the variability from different sources and under specific experimental settings variance components can be estimated, and (4) recognize that some amount of variability will always remain unexplained.

For this activity, we borrowed the approach of [Pillay et al. \(2008\)](#) who asked students to obtain

measurements of time and of distance. Specifically, we adapted the measurands to be the length of travel time by car and distance between Landmark 1 and Landmark 2. This exercise can be easily modified to incorporate two landmarks that students at a different university can relate to. The component that will be considered “parts” are the measurements of time and distance from a set of directions obtained from three different websites: MapQuest, Google Maps, and Yahoo!. Variability in the travel time and distance is reported by the different websites and students are asked to list at least five possible sources of measurement error that could have introduced this variability. Not all measurements are unique as multiple websites report identical measurements for travel time. The next question asks students if this implies that the repeated measurement value indeed corresponds to the true value of the measurand.

The second part of the activity presents students with multiple data collection scenarios. For each situation and given the available data, students must choose the most likely source of variability in the resulting measurements. For example, if one student repeated the trip three times, each time using a different set of directions, but always measuring the travel time with the same stopwatch, what variance components can be estimated? The assignment concludes with what we hope to be obvious after seeing the complete set of metrology material. The final question asks if it is possible to obtain, with complete certainty, an exact measurement equal to the true value of the measurand for the distance between the two landmarks followed by a separate question asking if this is possible for the corresponding travel time. Separating these questions with respect to the measurands is important because our pilot study confirmed findings by [Pillay, et al. \(2008\)](#), namely that many students think about measures of distance and measures of time differently.

5. Methodology

To evaluate the implementation of the metrology material we assess students’ factual and conceptual understanding of the presented content and surveyed students’ attitudes towards statistics. Both were given as a pre- and a post-evaluation, allowing us to observe changes in students’ understanding and attitudes after the presentation of the lecture and lab session material.

5.1 Understanding Component

The understanding assessment consists of twenty questions obtained from three already existing repositories: CAOS (Comprehensive Assessment of Outcomes for a first course in Statistics; [delMas et al. 2007](#); website: <https://apps3.cehd.umn.edu/artist/caos.html>), ARTIST (Assessment Resources Tools for Improving Statistical Thinking; website: <https://apps3.cehd.umn.edu/artist/index.html>), and a questionnaire developed by [Buffler et al. \(2009\)](#). Selected questions from the CAOS and ARTIST scales focus on the interpretation of variability both numerically and graphically addressing the aforementioned learning objectives B2, M2 and M4. Within the field of statistics we are not aware of an instrument that assesses students’ understanding of the basic concepts of metrology. For this reason, we refer to an existing assessment ([Buffler et al. 2009](#)) consisting of metrology related questions from the physics education literature. While we did not find that the instrument had been evaluated for content validity and reliability, the instrument has been used in published research studies. For

the set of selected questions, we made small modifications with respect to context and style while leaving the nature of the questions unchanged. Students were presented with similar examples during lecture and lab activities. Some of these questions pose a scenario, present the student with four statements, and ask them to choose which one they most closely agree with ([Buffler et al. 2009](#)). Other questions present the students with the results of an analysis and ask them to compare the sources of variability or to make decisions based on a comparison.

The content questions as they were originally presented to students in the pilot study are given in [Appendix A](#). Following the implementation of the pilot study and in response to reviewer feedback, we felt, however, that some edits to the questions will help clarify their meaning and lead to less ambiguity in the correct response. Possible improvements to specific questions are discussed immediately after they appear in [Appendix A](#). These modifications should be taken into consideration if the metrology material is to be implemented again. We discuss the results derived from students' responses to the assessment questions in Section 6.

5.2 Attitude Component

Because students often perceive statistics as irrelevant to both their personal as well as their professional life ([Snee 1993](#); [Schau and Emmioglu 2012](#)), one objective of embedding metrology into introductory statistics courses is to make statistics more attractive to undergraduate students by connecting their learning activities to real-world applications. We used a subset of nine questions from the SATS-36 (Survey on Attitudes Towards Statistics) by [Schau \(2003\)](#). The subset was selected to address opinions about the usefulness of statistics in students' lives ('value' dimension of the SATS scale) and students' perception of the difficulty of the material (two items addressing 'cognitive competence'; one item from the 'difficulty dimension'). These questions were selected based on what we considered most in-line with our study objectives. We realize that arguments can be made for a different set of questions. Seven of the questions are negatively worded and thus were reverse coded in the analysis so that higher responses correspond to a higher perception of the usefulness or understanding of statistics. Items included in the survey are:

- (1) (Reverse coded) Statistics is worthless. (value)
- (2) (Reverse coded) Statistical thinking is not applicable in my life outside my job. (value)
- (3) I use statistics in my everyday life. (value)
- (4) (Reverse coded) Statistics conclusions are rarely presented in everyday life. (value)
- (5) (Reverse coded) I will have no application for statistics in my profession. (value)
- (6) I can learn statistics. (cognitive competence)
- (7) (Reverse coded) Statistics is irrelevant in my life. (value)
- (8) (Reverse coded) I will find it difficult to understand statistical concepts. (cognitive competence)
- (9) (Reverse coded) Most people have to learn a new way of thinking to do statistics. (difficulty)

The classroom material and evaluation metrics were administered to two sections of an introductory statistics course at Iowa State University during the Spring 2011 semester. This course is designed for students in the liberal arts and sciences and provides an introduction to the basic statistical concepts including descriptive statistics, correlation and regression, confidence

intervals, and hypothesis testing. A 2-hour laboratory component is part of this course, in addition to three, 50-minute lectures per week. The metrology material was presented during two consecutive laboratory sessions. Each lecture took approximately 60 minutes to complete and was immediately followed by the corresponding group activity. Working in groups of size three to five, the students were able to complete the activity in 30-45 minutes. Courses that do not have a laboratory session can implement the activities during lecture, and, although group work is preferred, both activities may be used as homework assignments to be completed by individual students, making the material suitable for an online course as well.

Seventy-eight students completed the pre- and post-assessment on the understanding questions as well as the survey on attitudes toward statistics. All evaluations were administered online. Students were given a week before the first presentation and a week after its completion in the second lab period to finish the assessments. Therefore, there were from two weeks to at most a month between the two attempts. The solutions to the content questions were not made available after the first attempt. The participants were aware of the details of the study prior to any data collection and were given the option to opt-out at any time. The Institutional Review Board at Iowa State University approved the study protocol, consent form, and all participant-related assessment and survey materials.

6. Results

The change in students' responses on the pre and post content assessments and the pre and post attitudes surveys are examined here to describe the effect of the metrology material in the pilot study. Because all students belonged to the same classroom and were introduced to the material by the same instructor, inferential statements may not generalize to the entire population of introductory students or instructors. Thus, the results below are only descriptive in nature. We do not consider this a limitation of the study because it served primarily as a pilot study for the purpose of obtaining feedback on the value of the metrology material in regard to introducing students to the existence and importance of measurement error, increasing students' perception of the usefulness of statistics and how a discussion on sources of measurement variability affects students' understanding of variability. The questions on the content assessment have been divided into three subsets based on the broad learning objective(s) (B2 or B3) to be addressed. Of the twenty questions on the assessment, seven test the understanding of metrology or measurement error only (B3), six focus on the understanding of variability within the context of metrology (B2 and B3), and seven address the understanding of variability only (B2). The last seven were obtained from the ARTIST assessment instrument and CAOS test. Broad learning objective B1 is assessed in the attitudes survey.

6.1 Content Assessment Results

Out of the 78 students who completed both the pre- and post-content assessment, 59, or 75.6%, improved or performed equally well after the presentation of the metrology material. The median amount of improvement was about 10 percentage points, or equivalently two additionally correct answers on the post assessment compared to the pre assessment.

[Figure 1](#) displays the distributions of the differences in students' percentage of correct answers between the post and pre assessment for each subset of questions. Note that a difference greater than zero, plotted to the right of the thick vertical line, indicates an improvement in the score for that category. [Table 1](#) shows the percentages of students whose performance improved or stayed consistent within each subset of questions. The top plot of [Figure 1](#) corresponds to the difference in the percentage of correctly answered questions of the seven questions testing students' knowledge of metrology. About 77% of the students showed an improvement or no change in answering the metrology related questions. The longer right tail of the distribution indicates that the greatest individual student improvements were seen in this category of questions. The distribution of the differences in the scores for variability within metrology questions is shown in the second plot of [Figure 1](#). The majority of students showed improvement in this category with 64, or 82%, improving or maintaining the number of correctly answered questions between the pre and post assessments. Although still a majority, a somewhat lower percentage of students (69%) improved or remained the same in the third category regarding the variability only questions. Students showing improvement or consistency in this category improved by one additional correct question, on average.

Figure 1. Distributions of the differences in scores for each student on the content assessment with results grouped by learning objective of the question.

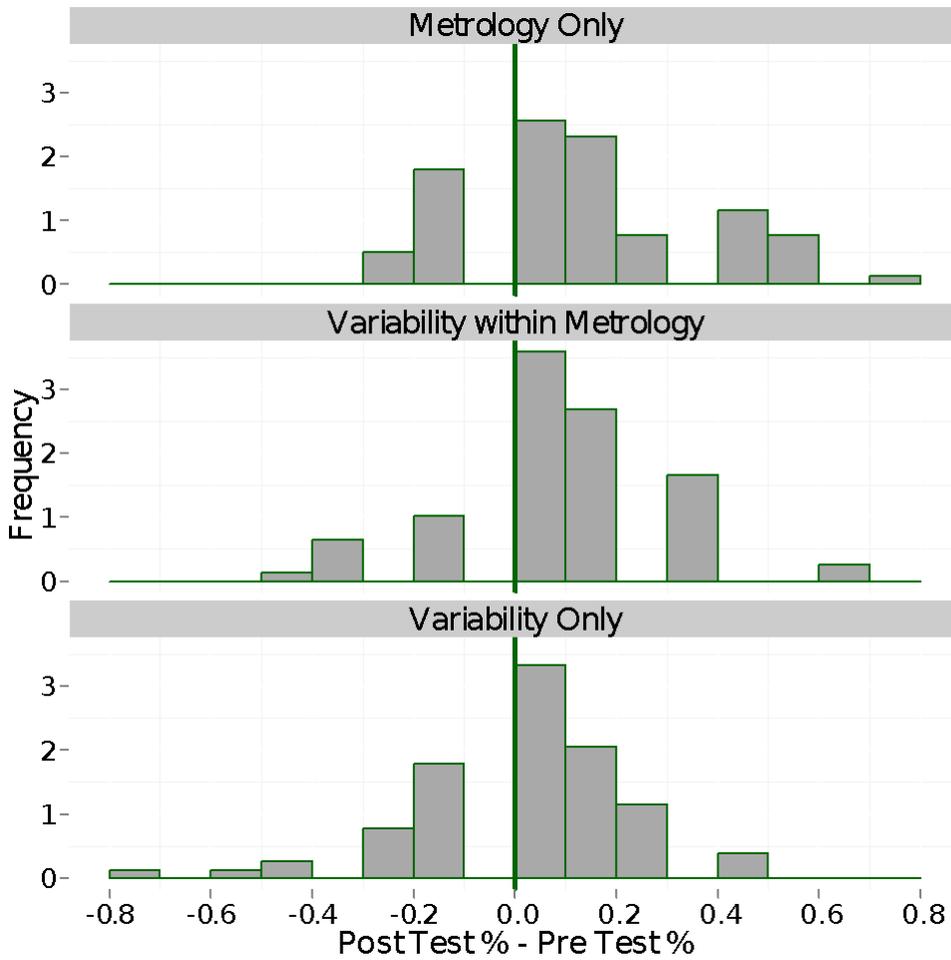


Table 1. Percentage of students whose score increased or stayed consistent for each subset of questions from the pre to post assessment.

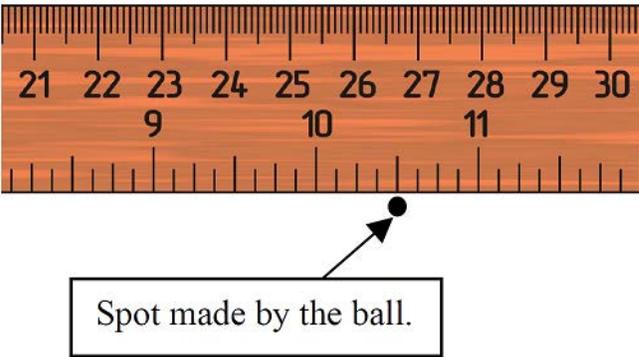
	% of students with equal or higher response on post assessment	% of students with higher response on post assessment
Metrology only	76.9	51.3
Variability within Metrology	82.1	46.2
Variability Only	69.2	35.9

6.1.1 Results for Metrology Questions

The first set of questions aimed to test students' understanding of concepts within the field of metrology. Relevant learning objectives are given in parentheses. Specifically, the questions asked students to read and interpret measurements (M1), to assess a measurement scenario with an emphasis on the importance of repeated measurements (M2), to recognize the omnipresence of measurement error (M3), and to recognize and account for measurement error even when reporting a single measurement (M1). More students improved on questions about reading and interpreting measurements than any other type of question. An example of this type of question as it was presented to the students is shown in [Figure 2](#) as well as a summary of students' responses on both the pre and post assessments in [Table 2](#). The number of students who correctly identified answer (d) increased from 6 to 28 after being presented with the metrology material. Important to note also is the change in the number of students who chose answer (a). This answer directly addresses learning objective M3 and represents a common misconception that the correct procedure for collecting measurements is to measure once and this will result in the correct value ([Deardorff 2001](#), p. 26). With the introduction of the material, the number of students who chose this answer decreased from 16 to 6.

Figure 2. Content assessment question to test students' ability to read and interpret measurements. This question attempts to evaluate learning objectives M1 and B3.

A group of students are working together on an experiment. The first task is to determine the distance a ball rolls down a slope from a determined height. The group uses a ruler to measure the distance. What they see is shown in the picture below.



Suppose the group consists of Student A, Student B, Student C, Student D, and Student E. Which student do you most closely agree with?

- (a) Student A who says: The distance is exactly 10.5 inches.
- (b) Student B who says: The distance is approximately 10.5 inches.
- (c) Student C who says: The distance is between 10.25 and 10.75 inches.
- (d) Student D who says: The distance is between 10.4 and 10.6 inches.
- (e) Student E who says: I don't agree with any of you.

Table 2. Results of question shown in [Figure 2](#).

		Post Assessment					Total
		(a)	(b)	(c)	(d)	(e)	
Pre Assessment	(a)	4	7	1	4	0	16
	(b)	1	32	1	18	0	52
	(c)	1	0	0	2	0	3
	(d)	0	1	1	4	0	6
	(e)	0	1	0	0	0	1
Total		6	39	3	28	0	78

6.1.2 Results for Variability within Metrology Questions

Six questions on the assessment examined students' ability to interpret variability within the context of metrology. Each of these questions presents students with a graphical display of repeated measurements and asks students to compare the variability between components of a source of variability, such as repeated measurements performed by three different operators. These questions are intended to test learning objective M4. An example of the type of question from this section is shown in [Figure 3](#). The 78 responses on both assessments to this question are shown in [Table 3](#). The number of correct responses increased from 21 to 33 or by 57%. However, students had a difficult time distinguishing between variance and range, as response (c) was the most selected response on both assessments and the correct answer to this question, (b), requires perhaps a more detailed understanding of the arithmetic behind variance calculations than was emphasized in the presented metrology materials.

Figure 3. Content assessment question to test students' ability to graphically compare variability. This question attempts to evaluate learning objectives M4, B2 and B3.

In an experiment three operators measure an object with the same measurement device exactly three times. A chart displaying graphically the measurement values obtained by the operators is shown below. Do any of the operators show higher variability than the other two operators?

Operator	Measurement 1	Measurement 2	Measurement 3
1	379.5	380	380.5
2	377.5	378	378.5
3	375.5	376	376.5

(a) Operator1, because this operator has the highest measurement values overall.
 (b) Operator 2, because two of the three measurements are more closely clumped than for any of the other 2 operators.
 (c) All of the operators have a similar amount of variability in their measurements as the distance between the smallest and the largest measurement is about the same for all operators.
 (d) This cannot be determined from this graph without further information.

Table 3. Results of question shown in [Figure 3](#).

		Post Assessment				Total
		(a)	(b)	(c)	(d)	
Pre Assessment	(a)	0	2	3	1	6
	(b)	0	13	7	1	21
	(c)	1	15	27	2	45
	(d)	0	3	3	0	6
Total		1	33	40	4	78

6.1.3 Results for Variability Only Questions

The last subset of questions on the content assessment focused on quantifying students' understanding of variability outside of a metrology context. The questions in this subset asked students to interpret a standard deviation in the context of a problem and again asked them to compare variability from two datasets given graphical displays of both. An example of this type of question is displayed in [Figure 4](#), and the student responses are shown in [Table 4](#). Many students were able to correctly respond even before the presentation of the metrology material, as 59 students initially chose (c). In the pilot study the pre assessment was given midway through the semester, thus students had recently been introduced to variability, which could account for the higher baseline knowledge of the concept. Only four of those students were unable to respond correctly the second time, while twice as many previously incorrect students chose the correct response.

Figure 4. Content assessment question to test students' ability to interpret variability outside the context of metrology.

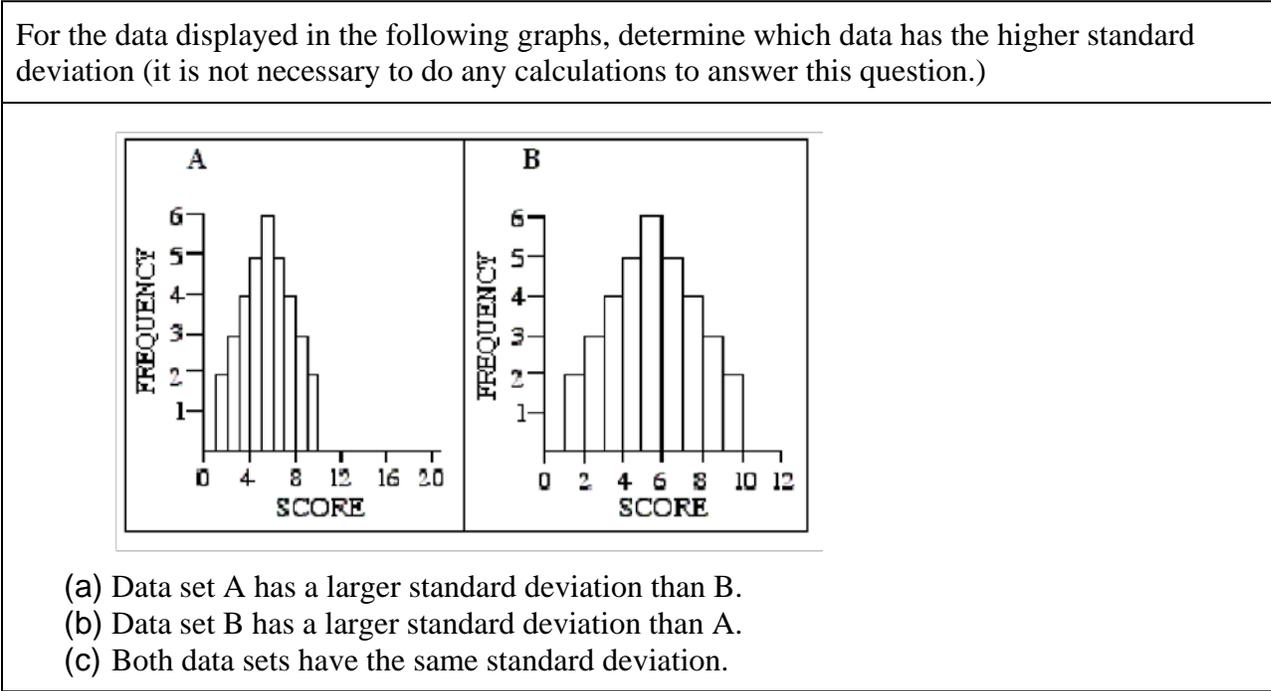


Table 4. Results of question shown in [Figure 4](#).

		Post Assessment			Total
		(a)	(b)	(c)	
Pre Assessment	(a)	0	0	3	3
	(b)	1	10	5	16
	(c)	1	3	55	59
Total		2	13	63	78

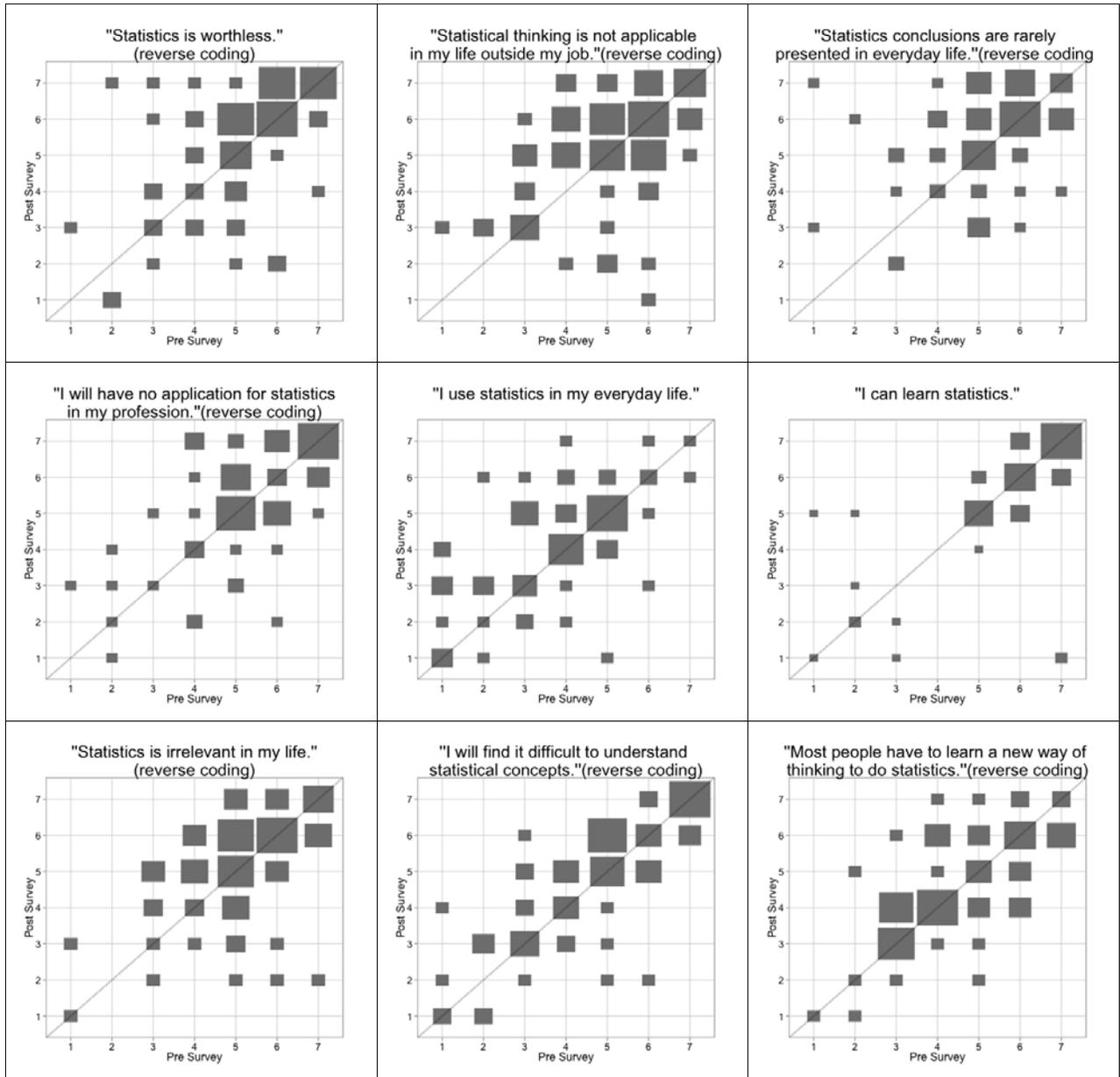
6.2 Attitudes Questionnaire Results

The attitudes questionnaire served as a measure of how students perceive the usefulness of statistics before and after the completion of the metrology lectures and group activities. For each survey, students were presented with the same nine statements and asked to state their level of agreement on a Likert-style scale of 1 to 7, where 7 represents the highest rating of perceived relevance or ability. The pre survey in the pilot study was administered midway into an

introductory statistics course and thus represents the students' perception and opinions of statistics at that time.

The results of both surveys for all nine statements are graphically summarized in [Figure 5](#) using fluctuation diagrams ([Wickham and Hofmann 2011](#)). A fluctuation diagram is a graphical display of a contingency table. Specifically, the horizontal axis shows a student's response on the pre-survey against the corresponding post survey response located on the vertical axis. The height and width of the plotting symbols are proportional to the square root of the count of the corresponding cell in the contingency table. Note that seven of the nine statements have been reverse coded, so an increase in perceived relevance or understanding from the pre to post survey is plotted above the diagonal line in each plot. Therefore, an increase in response for the statement "Statistics is worthless" indicates that the student perceives statistics as less worthless or more valuable after the presentation of the metrology material.

Figure 5. Pre and post survey responses plotted by student for each question as fluctuation diagrams. Those responses that have been reverse coded are indicated in the title. An increase in perception is plotted above the diagonal line in all plots.



An analysis of [Figure 5](#) shows that the boxes on or above the diagonal line have greater area than those below it indicating that most students did increase or stay consistent in their perception of the relevance and their ability to learn statistics. [Table 5](#) displays the percentages of students whose response on the post survey was equal or higher for each statement and the percentage of those whose response strictly increased, i.e., the percentage of students whose pair of responses fall on or above the diagonal line and the percentage above the line. The table indicates the largest number of increased or consistent responses were for the statement, “I use statistics in my

everyday life.” The plot that corresponds to this statement indicates that more students scored this unfavorably on the pre survey than any other statement. Therefore, the large percentage of increased responses is encouraging. The largest percentage of strictly larger responses was to the statement “Statistical thinking is not applicable in my life outside my job.” The statement with the lowest strict increase in perception was “I can learn statistics.” A view of the fluctuation plot for this statement indicates that one reason for the small increase is that this statement received favorable remarks on the pre survey. It should be remembered that the pre survey was taken after students had been through about half a semester of an introductory statistics course.

Table 5. Percentage of students whose perception of the usefulness and understandability of statistics increased or stayed consistent for each statement from the pre to post survey.

	% of students with equal or higher response on post survey	% of students with higher response on post survey
“Statistics is worthless.” (reverse coded)	76.9	41.0
“Statistical thinking is not applicable in my life outside my job.” (reverse coded)	73.1	43.6
“Statistical conclusions are rarely presented in everyday life.” (reverse coded)	75.6	39.7
“I will have no application for statistics in my profession.” (reverse coded)	73.1	30.8
“I use statistics in my everyday life.”	80.1	42.3
“I can learn statistics.”	78.2	15.4
“Statistics is irrelevant in my life.” (reverse coded)	67.9	35.9
“I will find it difficult to understand statistical concepts.” (reverse coded)	79.4	38.5
“Most people have to learn a new way of thinking to do statistics.” (reverse coded)	72.0	30.8

7. Conclusions and Discussion

Measurement is a critical component in all statistics courses ([Vardeman et al. 2010](#)). By incorporating metrology in the statistics curriculum, variability and measurement error may be illustrated through concrete, relatable examples. Further, students will be challenged to connect statistical concepts to metrology concepts such as controlling for sources of variability when

taking measurements as part of data collection, recognizing the importance of precision and bias in a measurement process, and understanding the implications of measurement error on statistical inference. This pilot study provided encouragement as our comparative study of the abilities and attitudes of introductory students who participated in this study indicated an improvement after the presentation of the metrology material. The students were better able to read and interpret measurements, understand variability and make comparisons based on that variability. Students' attitudes towards the usefulness of statistics also showed an increase. A more extensive study that allowed for statistical inference regarding the effectiveness of the metrology material can thus be considered for future research.

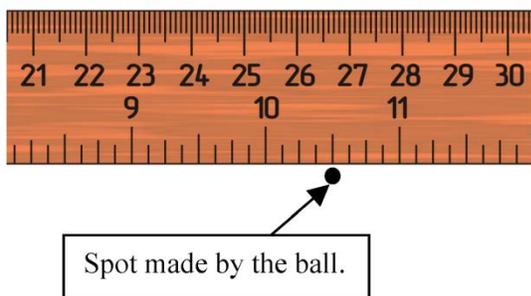
A less measurable benefit of the presentation of the metrology material was that students were forced to think about measurements in a new and possibly unfamiliar way. This is especially important for non-science students enrolled in an introductory level statistics course who may not be introduced to this concept in a lab science course. Metrology introduces students to the notion of questioning concepts that were once taken without uncertainty, which is what drives inquiry, good decision-making, and the field of statistics. We envision former students pumping gas into their car or weighing bananas at the grocery store checkout wondering about the measurement error of those devices and the consequences of that error for them.

Appendix A: Content Assessment

The following set of questions was used to measure students' understanding of the presented material. Correct answers are indicated with italicized text. The questions are listed below with the exact phrasing as it was presented to the students as part of the pilot study. After implementing the material and considering reviewer comments, possible ideas for revisions in future implementations are discussed immediately following some questions.

Question 1. A group of students are working together on an experiment. The first task is to determine the distance that a ball rolls down a slope from a determined height. The group uses a ruler to measure the distance. What they see is shown in the picture below.

Suppose the group consists of Student A, Student B, Student C, Student D, and Student E. Which student do you most closely agree with?



- (a) Student A who says: The distance is exactly 10.5 inches.
- (b) Student B who says: The distance is approximately 10.5 inches.
- (c) Student C who says: The distance is between 10.25 and 10.75 inches.
- (d) *Student D who says: The distance is between 10.4 and 10.6 inches.*
- (e) Student E who says: I don't agree with any of you.

Discussion of Question 1: The purpose of the question is to assess whether students can correctly report the distance from this particular roll, given that this is all the information that is available. This requires the students to report the measurement as an interval. Hence, an alternative formulation for the last sentence could be: "The students conduct the experiment once and what they see is shown in the picture below."

Question 2. The following discussion takes place between four of the students.

Student A: We should roll the ball a few more times from the same height and measure the distance each time. This will give us more information about the true measurement value even if we get distinct values every time.

Student B: Why, we measured the distance already and know the true value of the distance. Measuring more than once creates only uninformative variation in the data.

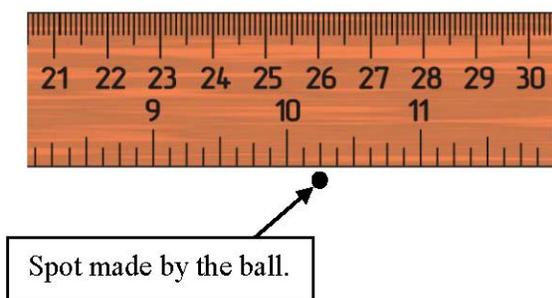
Student C: We should roll the ball down the slope one more time and average the two measurement values. This way we account for the measurement variability from the ruler.

Student D: We should roll the ball down the slope until we get the same measurement over and over again. This will be the true measurement value.

Which student do you most closely agree with?

- (a) Student A.
- (b) Student B.
- (c) Student C.
- (d) Student D.

Question 3. The group of students decides to roll the ball again from the same height as before. They use the same ruler to measure the distance of the second roll, and what they see is shown below. Which of the students do you mostly agree with?



- (a) Student A who says: Based on this measurement the distance is exactly 10.25 inches.
- (b) Student B who says: Based on this measurement the distance is approximately 10.25 inches.
- (c) Student C who says: Based on this measurement the distance is between 10.24 and 10.26 inches.
- (d) Student D who says: Based on this measurement the distance is between 10.125 and 10.375 inches.
- (e) Student E who says: I don't agree with any of you.

Discussion of Question 3: The purpose of this question is to assess whether students can correctly report the distance from this particular, single roll of the ball. The last sentence of the question could be modified to read, “If it were desired to report the result of this single roll, which of the students do you most agree with?”

Question 4. After two rolls from the same height, the students have the following readings:

Reading 1: distance = 10.5 inches

Reading 2: distance = 10.25 inches

Which student do you mostly agree with?

- (a) Student A: We know enough. We don't need to roll the ball again.
- (b) Student B: We need to roll the ball just one more time and average the three measurements.
- (c) *Student C: Three rolls will not be enough. We should roll the ball several more times and measure the distance each time.*
- (d) Student D: We should roll the ball again until we get the same measurement twice and record that as our distance.

Discussion of Question 4: The intent of the question is to assess students' understanding of misconceptions addressed in Section 2.2. This original question set-up does not provide students with sufficient information to distinguish between the correctness of the first three options. The question could potentially be improved by adding additional measurements and rewriting the question and answers to highlight a single misconception.

Question 5. The lecturer now comes around with a special new electronic device that has a digital display and that can be used to measure the distance the ball rolls down the slope. The electronic reading from this device is shown in the picture below.



Which student do you mostly agree with?

- (a) Student A who says: The distance is exactly 10.3 inches.
- (b) Student B who says: The distance is approximately 10.3 inches.
- (c) Student C who says: The distance is between 10 and 11 inches.
- (d) *Student D who says: The distance is between 10.2 and 10.4 inches.*
- (e) Student E who says: I don't agree with any of you.

Discussion of Question 5: Similar to the discussion of Questions 1 and 3, the wording of this question should reflect that the goal is for the students to report the results of this single roll.

Question 6. When the group is finished, two students discuss how they can improve their rolling ball experiment next time.

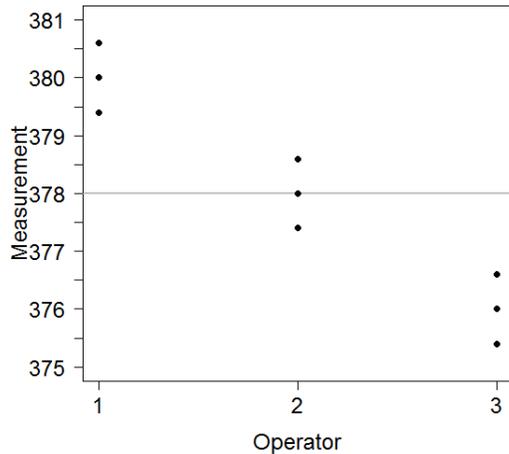
Which student do you mostly agree with?

- (a) Student A who says: If we practice enough and work very carefully, all our readings will be the same. Then we will know the true value of the distance.
- (b) *Student B who says: No, even if our readings are all the same, we will still not know the true value of the distance.*

Question 7. If an object is found to have a mass of 90.85 grams with a stated uncertainty of 0.05 grams, then the measurement result should be reported as
The mass of the object is between [90.80] and [90.90] grams.

Discussion of Question 7: Possible rewording to make more clear, “If an object is found to have a mass of 90.85 grams based on a single weighing on a balance with a stated measurement uncertainty of 0.05 grams, then the measurement result should be reported as:”

Question 8. In an experiment three operators measure an object with the same measurement device exactly three times. A chart displaying graphically the measurement values obtained by each operator is shown below. Do any of the operators show lower variability than the other two operators?

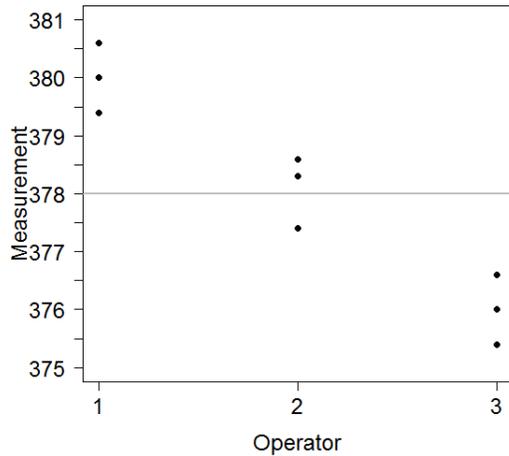


- (a) Operator 3, because this operator has the lowest measurement values overall.
- (b) *No, all of the operators have a very similar amount of variability in their measurements.*
- (c) This cannot be determined from this graph without further information.

Discussion of Questions 8-13: Questions 8 through 13 require students to determine differences in the variability of repeated measurements from three different operators (questions 8 through 11) or 10 different parts (questions 12 and 13). Because the sample size for each set of repeats is only 3, any hypothesis test would likely not result in significant differences regarding the actual variances. We propose two possible clarifications. If the intended objective of the questions is to focus on the observed amount of variability in the repeated samples, then this should be stated in the question set-up. For example, question 8 could be reworded as “Do any of the operators show less variability in the observed samples than the other two operators?” Further, the data could be plotted as boxplots instead of dotplots to de-emphasize the small sample sizes and to focus rather

on the range as a measure of variation. This clarification would be appropriate for the students in the pilot study and the time in the semester we introduced the metrology material. If, however, the instructor's objective is to assess if students understand that this set-up does not allow for clear-cut inferential statements about possible differences in the variances, then question 8 could be rewritten as "Based on the information displayed in the chart, do any of the operators show significantly less variability in the repeated measurements than the other two operators?" This would change the correct answer for question 8 from (b) to (c). Displaying the data with boxplots for this option would prevent students from attempting an actual hypothesis test and would emphasize conceptual understanding.

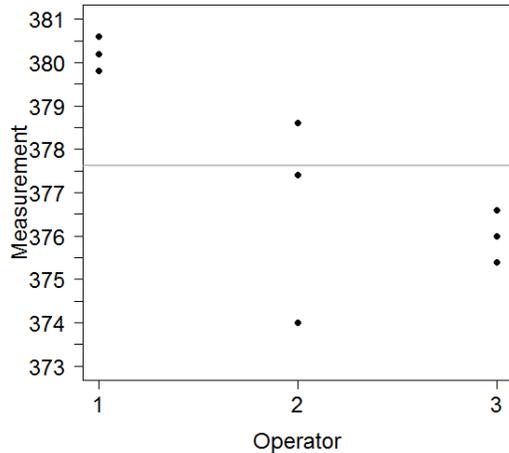
Question 9. In an experiment three operators measure an object with the same measurement device exactly three times. A chart displaying graphically the measurement values obtained by each operator is shown below. Do any of the operators show higher variability than the other two operators?



- (a) Operator 1, because this operator has the highest measurement values overall.
- (b) Operator 2, because two of the three measurements are more closely clumped than for any of the other 2 operators.
- (c) All of the operators have a similar amount of variability in their measurements as the distance between the smallest and the largest measurement is about the same for all operators.
- (d) This cannot be determined from this graph without further information.

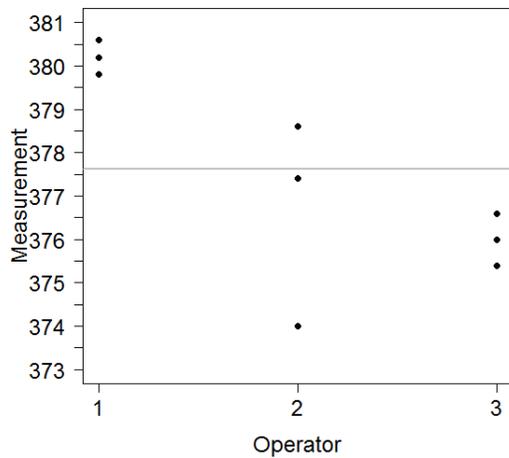
Discussion of Question 9: This particular question aims to evaluate students' conceptual understanding of variability through the means of graphically displayed data. Students in the pilot study struggled with this question and primarily focused on the range of the data for each operator. Further, the correct answer to this question requires perhaps a more detailed understanding of the arithmetic behind variance calculations than was emphasized in the presented metrology materials.

Question 10. In an experiment three operators measure an object with the same measurement device exactly three times. A chart displaying graphically the measurement values obtained by each operator is shown below. Do any of the operators show higher variability than the other two operators?



- (a) Operator 1, because all of the measurements are highest overall.
- (b) Operator 2, because these measurements are spread out over a larger range.
- (c) Operator 3, because this operators shows the smallest number of distinct measurement values.
- (d) This cannot be determined from this graph without further information.

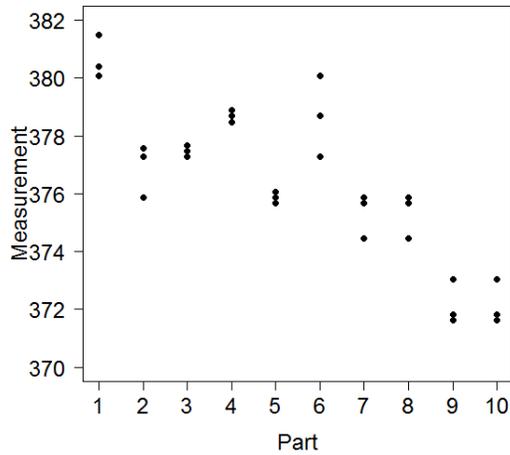
Question 11. In an experiment three operators measure an object with the same measurement device exactly three times. A chart displaying graphically the measurement values obtained by each operator is shown below. Do any of the operators show lower variability than the other two operators?



- (a) Operator 1, because the range of these measurements is smallest.

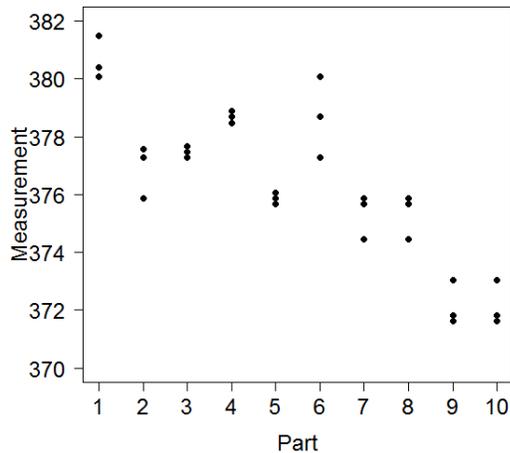
- (b) Operator 3, because these measurements are lower than all but one other measurement value.
- (c) This cannot be determined from this graph without further information.

Question 12. The experiment is changed and now one operator measures 10 objects (also called parts) with the same measurement device three times each. A chart displaying graphically the measurement values obtained by each part is shown below. Which part(s) show(s) the lowest amount of measurement variability?



- (a) Parts 9 and 10.
- (b) Part 5.
- (c) *Parts 3, 4 and 5.*
- (d) This cannot be determined from this graph without further information.

Question 13. The experiment is changed and now one operator measures 10 objects (also called parts) with the same measurement device three times each. A chart displaying graphically the measurement values obtained by each part is shown below. Which part(s) show(s) the highest amount of measurement variability?



- (a) Part 6.
- (b) Part 1.
- (c) This cannot be determined from this graph without further information.

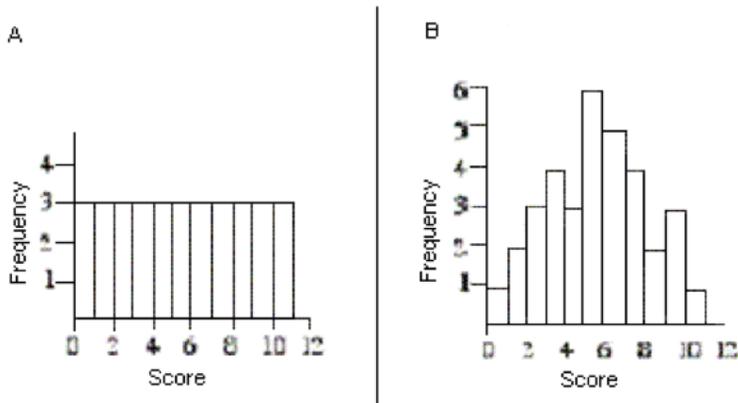
Question 14. A class of 30 introductory statistics students took a 15 item quiz, with each item worth 1 point. The standard deviation for the resulting score distribution is 0. You know that

- (a) About half the score were above the mean.
- (b) An arithmetic error must have been made.
- (c) *Everyone correctly answered the same number of items.*
- (d) The mean, median, and mode must all be 0.

Question 15. The 30 introductory statistics students took another quiz worth 30 points. On this quiz, the standard deviation was 1 point. Which of the following gives the most suitable interpretation?

- (a) All of the individual scores are one point apart.
- (b) The difference between the highest and lowest score is 1.
- (c) The difference between the upper and lower quartile is 1.
- (d) *A typical score is within 1 point of the mean.*

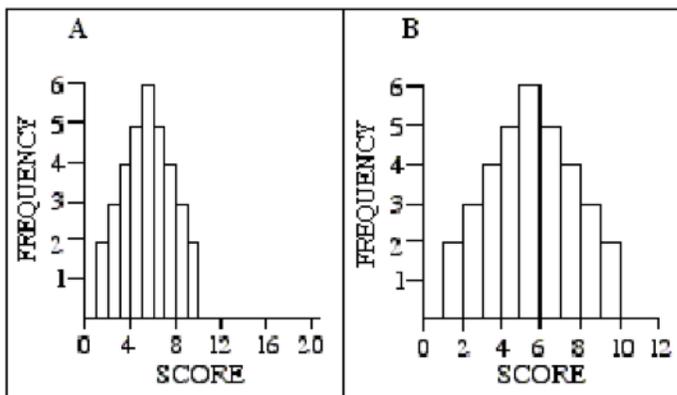
Question 16. For the data displayed in the following graphs, determine which data has the higher standard deviation (it is not necessary to do any calculations to answer this question.)



- (a) A has a larger standard deviation than B.
- (b) B has a larger standard deviation than A.
- (c) Both data sets have the same standard deviation.

Discussion of Questions 16-17: If the instructor chooses to highlight that the question focuses on the sample standard deviation, the question wording could be changed to ask "which data has the higher sample standard deviation?" However, the current wording is adapted from the ARTIST SCALE: Measures of Spread assessment.

Question 17. For the data displayed in the following graphs, determine which data have the higher standard deviation (it is not necessary to do any calculations to answer this question.)

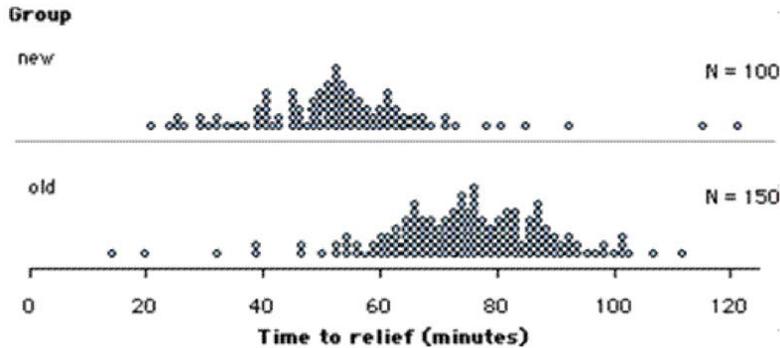


- (a) Data set A has a larger standard deviation than B.
- (b) Data set B has a larger standard deviation than A.
- (c) Both data sets have the same standard deviation.

Question 18. A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger

population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown in the graph below.

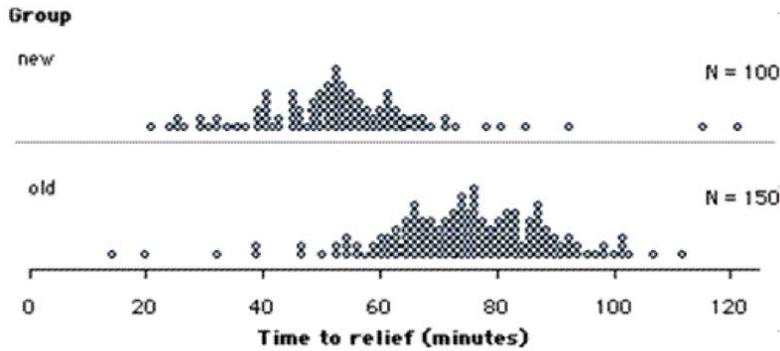
Based on this graph we can conclude that the old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.



- (a) True
- (b) False

Question 19. A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown in the graph below.

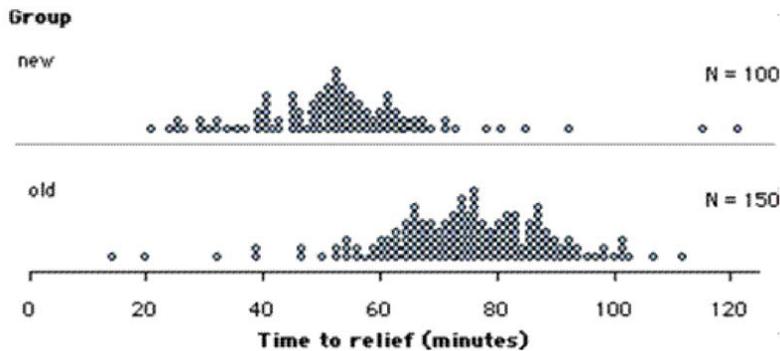
Based on this graph we can conclude that the average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.



- (a) True
- (b) False

Question 20. A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown in the graph below.

Based on this graph we would not conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.



- (a) True
- (b) False

Appendix B: Attitudes Questionnaire

The statements in this survey are designed to identify your attitudes about statistics. Each item has 7 possible responses. The responses range from 1 (strongly disagree) through 4 (neither disagree nor agree) to 7 (strongly agree). If you have no opinion, choose response 4. Please read each statement. Mark the one response that most clearly represents your degree of agreement or disagreement with that statement. Try not to think too deeply about each response. Record your answer and move quickly to the next item. Please respond to all of the statements.

1. Statistics is worthless.
2. Statistical thinking is not applicable in my life outside my job.
3. Statistics conclusions are rarely presented in everyday life.
4. I will have no application for statistics in my profession.
5. I use statistics in my everyday life.
6. I can learn statistics.
7. Statistics is irrelevant in my life.
8. I will find it difficult to understand statistical concepts.
9. Most people have to learn a new way of thinking to do statistics.

Acknowledgements

The authors would like to thank the Spring 2011 Principles of Statistics students for participating in this study. This work was supported in part by the Iowa State University Miller Faculty Fellowship Program (Genschel and Wilson). It was a creative component for a masters' degree at Iowa State University (Beyler née Borgen), was presented in the *Posters and Beyond* session at the 2011 USCOTS in Raleigh, NC (Casleton) and as a contributed paper in the Classroom Examples and Technique session at the 2011 JSM in Miami Beach, FL (Casleton). Lastly, the authors would like to thank two anonymous referees and two editors for insights and suggestions that helped improve the clarity of the manuscript, classroom materials, and assessments.

References

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

American Statistical Association (2005), "GAISE College Report," [online]. Available at www.amstat.org/education/gaise/GAISECollege.htm. Retrieved last on 31 May 2013.

Anderson, R.C., Shirey, L.L., Wilson, P.T., and Fielding, L.G. (1987), "Interestingness of children's reading material," In R.E. Snow and M.J. Farr (Eds.), *Aptitude, learning, and*

instruction. Vol. 3: Cognitive and affective process analyses, (pp. 287-299), Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

ARTIST, "Assessment Resource Tools for Improving Statistical Thinking," [online]. Available at <https://apps3.cehd.umn.edu/artist/index.html>. Retrieved last on 16 June 2013.

Buffler, A., Allie, S., and Lubben, F. (2001), "The development of first year physics students' ideas about measurements in terms of point and set paradigm," *International Journal of Science Education*, 23:11, 1137-1156.

Buffler, A., Lubben, F., Allie, S., and Campbell, B. (2009), "Introduction to Measurement in the Physics Laboratory," Version 3.5, [online]. Available at www.phy.uct.ac.za/people/buffer/labmanual.html. (Full questionnaires available at www.phy.uct.ac.za/people/buffer/edutools.html) Retrieved last on 31 May 2013.

CAOS, "Comprehensive Assessment of Outcomes in a First Statistics course," [online]. Available at <https://apps3.cehd.umn.edu/artist/caos.html>. Retrieved last on 16 June 2013.

CS Consultants, LLC. (2005), "Survey of Attitudes Towards Statistics (SATS)," [online]. Available at www.evaluationandstatistics.com/. Retrieved last on 31 May 2013.

Deardorff, D.L. (2001), "Introductory physics students' treatment of measurement uncertainty," PhD Thesis, North Carolina State University.

delMas, R., Garfield, J.B, Ooms, A., and Chance, B. (2007), Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28-58. [http://www.stat.auckland.ac.nz/~iase/serj/SERJ6\(2\)_delMas.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf)
[http://www.stat.auckland.ac.nz/~iase/serj/SERJ6\(2\)_delMas.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf)

delMas, R. and Liu, Y. (2005), "Exploring students' conceptions of the standard deviation," *Statistics Education Research Journal*, 4(1), 55-82.

Gal, I. and Ginsburg, L. (1994), "The Role of Beliefs and Attitudes in Learning Statistics: Towards an Assessment Framework," *Journal of Statistics Education*, 2(2). Available at <http://www.amstat.org/publications/jse/v2n2/gal.html>. Retrieved last on 01/31/2014.

Garfield, J.B. and Ben-Zvi, D. (2005), "A Framework for Teaching and Assessing Reasoning About Variability," *Statistics Education Research Journal*, 4, 92-99.

Garfield, J.B. and Gal, I. (1999), "Assessment and Statistics Education: Current Challenges and Directions," *International Statistical Review*, 67, 1-12.

Hammerman, J.K and Rubin, A. (2004), "Strategies for managing statistical complexity with new software tools," *Statistics Education Research Journal*, 3(2), 17-41.

JMP, Version 10. SAS Institute Inc., Cary, NC, 1989-2012.

Keeley, J., Zayac, R., and Correia, C. (2008), "Curvilinear Relationships between Statistics Anxiety and Performance among Undergraduate Students: Evidence for Optimal Anxiety," *Statistics Education Research Journal*, 7(1), 4-15.

Kimbrough, D.R. and Meglen, R.R. (1994), "A Simple Laboratory Experiment Using Popcorn to Illustrate Measurement Error," *Journal of Chemical Education*, 71, 519-520.

Lee, H.K. (2007), "Chocolate Chip Cookies as a Teaching Aid," *The American Statistician*, 61, 351-355.

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

Makar, K. and Confrey, J. (2005). "Variation-talk: Articulating meaning in statistics," *Statistics Education Research Journal*, 4(1), 27-54.

Moore, D. (1990), "Uncertainty," In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy*, (pp. 95-137), Washington, D.C.: National Academic Press.

Pacer, R.A. (2000), "How Can an Instructor Best Introduce the Topic of Significant Figures to Students Unfamiliar with the Concept?" *Journal of Chemical Education*, 77, 1435-1438.

Peters, S. (2011), "Robust understanding of statistical variation," *Statistics Education Research Journal*, 10(1), 52-88.

Pillay, S., Buffler, A., Lubben, F., and Allie, S. (2008), "Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory," *European Journal of Physics*, 29, 647-659.

Prilliman, S.G. (2012), "An inquiry-based density laboratory for teaching experimental error," *Journal of Chemical Education*, 89(10), 1305-1307.

Reading, C. (2004), "Student description of variation while working with weather data," *Statistics Education Research Journal*, 3(2), 84-105.

Reading, C. and Shaughnessy, M. (2004), "Reasoning about Variation," In D. Ben-Zvi and J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking*, (pp. 201-226), Dordrecht, The Netherlands: Kluwer Academic Publishers.

Reid, J. and Reading, C. (2008), "Measuring the development of students' consideration of variation," *Statistics Education Research Journal*, 7(1), 40-59.

Schau, C. (2003), Survey of Attitudes Toward Statistics (SATS-36). [Online: <http://evaluationandstatistics.com/>]

Schau, C. and Emmioglu, E. (2012), "Do introductory statistics courses in the United States improve students' attitudes?," *Statistics Education Research Journal*, 11(2), 86-94.

Schiefele, U. (1991), "Interest, learning and motivation," *Educational Psychologist*, 26(3&4), 299-323.

Séré, M.-G., Journeaux, R., and Larcher, C. (1993), "Learning the statistical analysis of measurement errors," *International Journal of Science Education*, 15(4), 427-438.

Snee, R. (1993), "What's Missing in Statistical Education?" *The American Statistician*, 47(2), 149-154.

Terezakis, E.G. (2010), "Introduction to Measurements and Basic Laboratory Techniques," [online]. Available at http://faculty.ccri.edu/eterezakis/1020_Exp_1_Intro_to_Measurements_and_Basic_Laboratory_Techniques.pdf. Retrieved last on 31 May 2013.

Vardeman, S.B., Wendelberger, J.R., Burr, T., Hamada, M.S., Moore, L.M., Jobe, J.M., Morris, M.D., and Wu, H. (2010), "Elementary Statistical Methods and Measurement Error," *The American Statistician*, 64, 46-51.

Weber, K., Fornash, B., Corrigan, M., and Neupauer, N.C. (2003), "The effect of interest on recall," *Communication Research Reports*, 20(2), 116-123.

Wickham, H. and Hofmann, H. (2011), "Product Plots," *IEEE Transactions on Visualization and Computer Graphics*, 17, 2223-2230.

Wild, C.J. and Pfannkuch, M. (1999), "Statistical thinking in empirical enquiry (with discussion)," *International Statistical Review*, 67(3), 223-265.

Zipp, A.P. (1992), "A Simple But Effective Demonstration for Illustrating Significant Figure Rules When Making Measurements and Doing Calculations," *Journal of Chemical Education*, 69, 291.

Emily Casleton & Ulrike Genschel
Department of Statistics & Statistical Laboratory
Snedecor Hall
Ames, IA 50011-1210
casleton@iastate.edu, ulrike@iastate.edu

Amy Beyler
UnitedHealth Group
Minnetonka, MN 55343
amy_beyler@uhc.com

Alyson Wilson
Department of Statistics
SAS Hall
Raleigh, NC 27695-8203
agwilso2@ncsu.edu

[Volume 22 \(2014\)](#) | [Archive](#) | [Index](#) | [Data Archive](#) | [Resources](#) | [Editorial Board](#) | [Guidelines for Authors](#) | [Guidelines for Data Contributors](#) | [Guidelines for Readers/Data Users](#) | [Home Page](#) | [Contact JSE](#) | [ASA Publications](#)