Calibrating the Difficulty of an Assessment Tool: The Blooming of a Statistics Examination

Bruce Dunham
Gaitri Yapa
Eugenia Yu
University of British Columbia

Journal of Statistics Education Volume 23, Number 3 (2015),
www.amstat.org/publications/jse/v23n3/dunham.pdf

Key Words: Statistics education; Bloom's taxonomy; Assessment; Calibrating difficulty

Abstract

Bloom's taxonomy is proposed as a tool by which to assess the level of complexity of assessment tasks in statistics. Guidelines are provided for how to locate tasks at each level of the taxonomy, along with descriptions and examples of suggested test questions. Through the "Blooming" of an examination – that is, locating its constituent parts on Bloom's taxonomy - the difficulty level of an examination paper in statistics can be pseudo-objectively assessed, via both its Bloom's Index and the proportion of marks allocated to higher order cognitive skills. One suggested application of the approach is in assessing the impact on student learning due to course transformations implemented incrementally over time. Assessment tools, in particular examination papers, can be compared for difficulty and student performance. A case study is provided in which examinations from an introductory course are Bloomed post-hoc and compared to student performances.

1. Introduction

Assessment has understandably been the focus of much research in statistical education. If, as Hubbard (1997) suggests, "Assessment drives the whole learning process," how students are assessed on a statistics course will be integral to how, and how much, they learn. Assessment serves purposes besides evaluating student learning, such as improving learning, providing instructor feedback, and reporting (Garfield 1994, Begg 1997). Numerous modes of assessment have been described in the literature; Garfield and Chance (2000), for example, list twelve approaches to assessment that have been adopted in statistical education, including the familiar (quizzes, examinations, group projects) to the less traditional (concept maps, critiques). Evidently, some modes of assessment will be better suited than others at probing student learning at different levels of mastery.

That there are different levels of mastery, even of introductory statistical concepts, seems to be broadly accepted. A primary aim, as Chance (2002) points out, is to "assess what you value" and there is consensus that "shallow," or "surface" learning is not the goal (see, for example, Steinhorst and Keeler 1995, Hubbard 1997, Schau and Mattern 1997, Garfield and Chance 2000). For example, the ability to perform a routine calculation is usually not considered a useful objective unless the learner can demonstrate the ability to interpret and transfer the procedure applied. Although there is agreement that assessing higher levels of mastery of statistics is important, particularly in the area of statistical inference (see for example, Alacaci 2004, and Lane-Getaz 2013), how one quantifies an assessment item in terms of the level of mastery it requires has not been explicitly addressed. In the context of a statistics course, we aim to illustrate how an assessment tool, such as an examination, can be objectively positioned in terms of the complexity of the tasks it contains.

Aligning teaching materials and curricula with assessment items at varying levels of complexity has been discussed by Marriott, Davies, and Gibson (2009) and Garfield et al. (2011). The latter suggest a two-way table (or "instructional blueprint") with rows for content topics and columns for cognitive demand. How to quantify cognitive demand is not described in detail by Garfield et al., although mention is made of the use of Bloom's taxonomy in this context. We pursue that suggestion in this work.

Bloom et al. (1956) proposed a taxonomy in an attempt to categorise levels of mastery in a general framework. Bloom identified a hierarchy of six levels of cognitive domains, ranging from mere recall up to synthesis and evaluation. Several uses have been proposed for the taxonomy, including informing course design and the development of formative assessment tools for students (see, for instance, Allen and Tanner 2002, 2007). Attempts have been made to use Bloom's taxonomy to calibrate the level of difficulty of an assessment tool, such as an examination (for instance Crowe, Dirks, and Wenderoth 2008, in biology, and Zheng, Lawhorn, Lumley, and Freeman 2008, for pre-med tests). We extend this field of research to the context of statistics. Overall, our work attempts to address the following three aims:

1. Interpreting Bloom's taxonomy in the context of statistical education.
2. Proposing how to locate on Bloom's taxonomy typical assessment tasks assigned in undergraduate statistics courses, particularly at the introductory level, leading to a pseudo-objective numerical measure of the complexity level of an assessment tool.
3. Suggesting ways in which the "Blooming" process can be used in practice in statistical education.

A possible application of the work here is in the context of course transformation. Following the Guidelines for Assessment and Instruction in Statistics Education (GAISE) (Aliaga et al. 2012), many instructors in statistics are adopting alternative pedagogy in their courses, such as incorporating group-based activities, clickers questions, on-line homework tools, and novel uses of technology. Evaluating the effectiveness of such interventions can be difficult, largely due to confounding. For instance, Prince (2004) summarises some of the pitfalls and problems in determining the effectiveness of active learning methods, and Zieffler et al. (2008) critique various studies in the context of statistics education. In summary, since transformations to a course can occur incrementally, over time, and since it may be practically impossible to perform true experiments as in the testing of a new drug treatment via a clinical trial, convincing evidence about the benefits to student learning due to course transformation can be elusive. In addition, producing the kinds of evidence typically encountered in studies, such as gains on scores on concept inventories or matched examination questions, is both time-consuming and entwined with ethical issues. If, with modest effort, an objective measure can be provided of the difficulty level of a test, then comparisons can be made not only across different tests, but also based on how students' performance relates to observed difficulty.

The authors faced the task of measuring the effect on student learning of a variety of changes to pedagogy in an introductory course, with the changes taking place over a period of years. The method described here appears a practical approach to assessing the impact of the changes on student learning and assessment. This application is described as a case study, illustrating a use of the "Blooming" process introduced.

The section that follows gives a brief introduction to Bloom's taxonomy, the revised form that will be adopted here, and its past use in calibrating difficulty levels of test questions. The subsequent section discusses Bloom's taxonomy applied to statistical education, with particular reference to the skills that are encountered in an introductory course. Section 4 describes details of a case study where the authors "Bloomed" examinations from an introductory course at their institution, summarising the results and including student performance data. Some discussion and conclusions follow in the Appendix, with examples of questions on various topics at different levels of the taxonomy. Such questions could be used to populate an instructional blueprint table of the kind suggested by Garfield et al. (2011).

2. Bloom’s taxonomy and calibrating assessment difficulty

The taxonomy proposed by Bloom et al. (1956) comprises six levels of mastery. The original labels are listed here with selections from the alternative terms suggested by the authors of the taxonomy: knowledge (define, duplicate, list, memorize), comprehension (classify, describe, explain, identify), application (apply, demonstrate, employ, illustrate), analysis (analyze, appraise, categorize), synthesis (arrange, compose, construct, create, design), and evaluation (appraise, argue, assess, compare). In part due to it not apparently replacing any existing framework for categorising cognitive dimensions, it was some years before the taxonomy gained much foothold in the education community. Once accepted, the structure sparked copious research. The taxonomy’s authors proposed two main goals for their work: educational (in particular, the taxonomy was suggested as a means to facilitate communication between educators) and psychological (such as how the taxonomy links to theories of how people learn). Primarily, the educational issues are of concern here.

As an attempt to categorise the behaviours of learners when performing various tasks, the taxonomy was designed to be discipline independent. It was appreciated by Bloom and his co-researchers that prior learning may shift the level of an activity on the scale, so that what is one level for an “expert” may be another level for a relative novice. Nevertheless, levels 1 to 4 (or 3, Madaus, Woods, and Nuttall 1973) are considered a hierarchy, in that higher levels cannot be achieved without applying lower ones.

The taxonomy aims to rank the *complexity* of a task, measuring depth of mastery, and as such does not correlate exactly with difficulty. For instance, asking a student to repeat the third bullet point on the second slide in the fifth lecture would demand great powers of recall from the student, but no demonstration of higher mastery. The first two levels of the taxonomy comprise "Low order cognitive skills" (LOCS), while levels 3 and above are usually deemed “High order cognitive skills” (HOCS). Tasks that require higher order skills should require a greater level of mastery, and there is some evidence that students tend to perform poorer on test questions requiring HOCS than those needing only LOCS (for instance, Zoller 1993, Knecht 2001, and Bailin 2002).

In the years that followed the introduction of the taxonomy, various researchers pointed to problems and limitations, both from an educational (Stanley and Bolton 1957, and Fairbrother 1975, for example) and a psychological perspective (such as Kropp and Stoker 1966, Madaus et al. 1973). On the one hand, educators appeared to have difficulty concurring when attempting to align tasks with levels on the taxonomy; on the other hand, there were issues with the hierarchical nature of the taxonomy representing realistic cognitive processes and other concerns about an agreement with accepted psychological models of learning. Other taxonomies were proposed, such as those of Ebel (1965) and Merrill (1971), but none rose to Bloom’s prominence among education academics, at least, despite evidence about its possible shortcomings.

A revised version of Bloom’s taxonomy (Anderson et al. 2001) addressed some of the potential failings of the original, and it is this modified categorization we follow here. In essence, the revised taxonomy aims to represent the two dimensions of both learning and cognition, in attempting to describe better the processes underlying tasks at various levels of complexity.

The revised taxonomy of Anderson *et al.* assigns verbs to cognitive levels, as in Table 1:

Table 1. Cognitive process levels with example descriptive verbs.

1. Remembering	Retrieving, recognizing, and recalling relevant knowledge from long-term memory; find out, learn terms, facts, methods, procedures, concepts.
2. Understanding	Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining. Understand uses and implications of terms, facts, methods, procedures, concepts.
3. Applying	Carrying out or using a procedure through executing, or implementing; make use of, apply practice theory, solve problems, use information in new situations.
4. Analyzing	Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing; take concepts apart, break them down, analyze structure, recognize assumptions and poor logic, evaluate relevancy.
5. Evaluating	Making judgments based on criteria and standards through checking and critiquing; set standards, judge using standards, evidence, rubrics, accept or reject on basis of criteria.
6. Creating	Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing; put things together; bring together various parts; write theme, present speech, plan an experiment, put information together in a new and creative way.

Anderson et al. (2001, Ch.16) weaken the purported hierarchy of the original taxonomy, noting, for example, that a task at level 2 (understanding) could be more cognitively demanding than a task at level 3 (applying), contending though that the “centre-points” of the scale do progress from most simple to most complex. There is still some debate about this point and the number of levels distinguishable for practical purposes. Madaus et al. (1973) suggest a "Y-shaped" structure, for example, with the first three levels as the stem, and with "Analysis" on one branch and the other two levels on the other, whereas Marzano (2001) posits just four levels in the cognitive domain, in effect subsuming Bloom’s levels 3-6 into two, “Analyzing” and “Knowledge Utilization.” For our goal here, assessing the complexity of the cognitive processes required to complete statistics test questions, we see no point in retaining any hierarchy in the two highest levels. Hence we combine levels 5 and 6 into one in what follows.

Although widely used in school education and in particular the training of teachers, until quite recently, neither Bloom’s taxonomy nor its revised version received much attention from educators in most disciplines in higher education. There have been recent attempts to use the taxonomy to enhance the learning of HOCS in an undergraduate biology class (Bissell and Lemons 2006), to create a tool to help biology educators better align assessments with teaching activities to improve the study and metacognitive skills of students (Crowe et al. 2008), and to characterize examinations in introductory calculus classes (Tallman and Carlson 2012). Our case study in section 4 more closely follows that of Freeman, Haak, and Wenderoth (2011), a study involving the “Blooming” of test papers as part of an appraisal of the effectiveness of instructional changes to an introductory biology course. In this study, three raters were used to allocate a level on the taxonomy to each question set on tests in a course over six different terms. Rules proposed by Zheng et al. (2008) were adopted to handle discrepancies between raters. For each test, a so-called Bloom's Index (BI) was computed, this being a weighted sum of the points per question, the weight being the assigned Bloom's level, and defined as

$$BI = 100 \times \frac{\sum_{i=1}^{N} P_i B_i}{5T}$$

where N is the number of parts on the test, P_i is the points available for part i, B_i is the level of that part (for $i=1,\ldots,N$), and T is the maximum score on test. (The divisor in the above is taken as $5T$, and not $6T$ as in Freeman et al. (2011), since it is taken here that there is no hierarchy in the two utmost levels.) Note that, for instance, an examination containing questions requiring only LOCS could not attain a Bloom's Index above 40. An examination consisting entirely of questions at the lowest level (recall) would have a Bloom's Index of 20.

The principle difficulty in applying the approach of Freeman et al. (2011) is equipping the raters to consistently assess the level of questions to a high level of concurrency. For this, raters must have a clear picture as to what a wide range of cognitive processes resemble in their discipline, and even experienced educators appear to struggle with this. Indeed, when summarising research appraising inter-rater reliability in this context, Seddon (1978, p.304) concluded "... the extent of perfect agreement decreases as the number of judges increases."

Reasons why raters often disagree when Blooming test questions have been identified. Raters may neglect to consider higher order skills when assessing the level of a question, stopping at lower level categories (Bissell and Lemons 2006, Crowe et al. 2008, and Zheng et al. 2008). There can be tendencies to focus on the perceived difficulty level of the topic assessed, rather than the cognitive process required (Lemons and Lemons 2013). Raters may also overlook information that had previously been provided to students, or the rating process may fail to provide necessary details about what students had been given (Crowe et al. 2008). All of these issues, and others, can lead to experienced instructors differing when assigning levels to questions on topics they have taught.

Casagrand and Semsar (2013, personal communication, May 3, 2013) have attempted to address the problem of rater reliability in the context of Blooming tests, creating a rubric they name the Bloom Dichotomous Key (BDK). Their key aim is to formalise the Blooming process, particularly for questions in physiology, and builds on Table 2 in Crowe et al. (2008). The approach appears promising, and some aspects are adopted here. For instance, it is important that the rater has information about what students were given, noting that any question is at level 1 (remembering) if the students have seen it before. Hence either the rater must be given such information, or the rater's responses must be modified in the light of such knowledge. Other aspects of the BDK rubric are less well suited to Blooming exams in statistics. For instance, asking the rater whether the students are being asked to "interpret data" would likely not be helpful in the context of Blooming a statistics examination question, although aspects of the follow-on questions (e.g., Are students re-describing the data to demonstrate they understand what the data represent? (level 2), Are students using the data to calculate the value of a variable? (level 3), Are students coming to a conclusion about what the data mean … and/or having to decide what data are important to solve the problem? (level 4)) have been incorporated into our process.

As Crowe et al. (2008, p.369) point out, "… each discipline must define the original classifications within the context of their field." This is not an exact science, and no rubric for raters of test questions in higher education disciplines is likely to produce complete concurrency across different raters. When used in the rating of 349 test questions, the BDK, for instance, resulted in total agreement between three raters only on around 40% of questions. This was, however, double the concurrency rate observed without the BDK (Casagrand and Semsar, personal communication, May 3, 2013).

3. Blooming in statistical education

The main challenge is interpreting the levels of Bloom's taxonomy in the context of the cognitive processes required to complete test questions in probability and statistics. This task proved more difficult than the authors had originally envisaged, even for questions in an introductory course, as simple heuristics proved inapplicable generally across the wide range of topics encountered.

The goal was to create guidelines that would be useful in Blooming tasks at all levels of instruction in statistics, from early school years up to graduate degree level. As a first step, the authors attempted to locate the definitions of Bloom's levels in the broad context of the discipline. These general descriptors are as follows:

1. Knowledge/Remembering: Recall, memorize, re-tell, repeat a definition, repeat a previously seen example, recall or identify a formula.
2. Comprehend/Understand: Put a concept into own words, identify an example of something, recognize a definition in an alternative wording, describe the key features.
3. Apply: Use a previously seen method to compute a value, create a graphic, or draw a generic conclusion from data.
4. Analyse: Formulate hypotheses and conclusions in the context of a study, deduce which known method is appropriate in a given scenario, recognise important structure in data.

5a. Create: Design a study to investigate a given hypothesis, propose a (to the student) novel solution to a problem.

5b. Synthesise/Evaluate: Collate and interpret information from multiple sources, compare and contrast alternative approaches, critique methodology.

Although the above descriptors give a flavour of how a level can be interpreted in the discipline, the terminology is too vague to be of much practical use. It is more helpful for raters to have descriptions of levels related to specific topics, and also see examples of questions at different levels on a topic. In what follows, we provide such information in the context of an introductory statistics course. This context seemed most apt, since a motivation was to assess examinations in an introductory course at our institution. As will be seen in section 4, the guidelines can readily be applied to higher level courses.

To best exemplify the levels, example topics from an introductory course were chosen as follows: exploratory data analysis (EDA), design, probability, sampling distributions, inferential methods, regression, and ANOVA. These topics are not mutually exclusive, but do serve to permit illustration of thought processes at different levels, within a context. Descriptors are provided for the levels for these topics in the following subsections, and selected example questions discussed. A more comprehensive set of example questions at each level is provided in the Appendix.

3.1 Knowledge/Remembering

Testing pure recall, questions at this level are quite rare on undergraduate statistics examinations at the authors' institution. Such questions target recall, without any attempt to gauge whether the learner attributes meaning to what is remembered. That said, any question which has been previously seen by a student must be at this level, since, however difficult, the student could feasibly simply have memorised the solution.

It is unclear how often questions at the lowest level of the taxonomy appear on university-level tests in statistics. Schau and Mattern (1997, p.5) suggested "A test often consists of a collection of items measuring primarily recall," though the comment may have referred to K-12 education. They cite an example (p.7) where the student is asked to identify the notation commonly used for the standard deviation, a question that would reasonably be placed at level 1, though the source of the question is not given.

Table 2, below, describes the types of tasks that one would perform at level 1 for the seven identified topics.

Table 2. Sample tasks for level 1.

EDA	Recall definitions, recall formula, identify graphic type, label parts of a graph.
Design	Recall definitions relating to experiments and observational studies.
Probability	Recall definitions and formulae.
Sampling distributions	Recall Central Limit Theorem (CLT), define a parameter and a statistic.
Inferential methods	Recall formula for a test statistic, remember definitions for hypothesis tests.
Regression	Recall definitions, identify response and predictor variables.
ANOVA	Recall definitions and formulae.

Two examples of questions at this level are provided here:

> Example 1.1: (EDA) Write down the formula for the variance of the numbers x_1, $x_2,\ldots,x_n$.
>
> Example 1.2: (Regression) The slope of a regression line and the correlation are similar in the sense that (choose all that apply):
> (a) they both have the same sign.
> (b) they do not depend on the units of measurement of the data.
> (c) they both fall between -1 and 1 inclusive.

(d) neither of them can be affected by outliers.
(e) both can be used for prediction.

The first example asks only for the learner to repeat a formula they have previously seen. In the second example, it is assumed that the learner has been told the underlying facts, and so merely needs to recall which statements are true. In the event that the student had not been told certain facts but had to deduce them, example 1.2 is no longer at level 1.

3.2 Comprehend/Understand

Educators often have difficulty with the term "understand," since in some sense the word encompasses entirely what we want students to achieve. In the context of Bloom's however, the term refers to there being meaning attached to something that is recalled. This could be demonstrated by being able to put a definition or explanation into novel wording, or identifying an example of something. So while, for instance, a student may be able to write out the formula for a sample variance (level 1), explaining the formula requires a deeper level of cognition.

Table 3 below describes some tasks representative of work at level 2.

Table 3. Sample tasks for level 2

EDA	Explain a statistic in words, describe the features of a graphic.
Design	Identify a study as an experiment or observational, describe features of an experiment using terminology.
Probability	Explain concepts in words, discuss an example of independence.
Sampling distributions	Explain CLT in words, identify parameters and statistics.
Inferential methods	Explain concepts in own words, identify which test has been applied given context and adequate information.
Regression	Explain least squares in words, describe regression line in simple terms.
ANOVA	Explain sums of squares (SS) in words, identify parameters, describe simple properties of SS.

It is suggested that level 2 can be the most difficult level to assign unambiguously. Note, for instance, that the ability to identify a study as an experiment is deemed to be at this level, and not level 1 since the task does not involve pure recall (unless the student has seen the study earlier). In practice, students are asked to do something to demonstrate understanding, and in our context are not typically asked to re-write definitions in their own words. Hence, distinguishing what tasks meet the level here can be open to interpretation, and some overlap with levels 3 and 4 seem inevitable. Two examples follow:

> Example 2.1: (EDA) For each of the following variables, indicate in the corresponding box a C if the variable is a categorical variable and a Q if it is quantitative.
> (a) Eye colour.
> (b) The cost of a car.
> (c) The number of bees in a beehive.

(d) The time between calls to call centre.
(e) The religions of students on a course.

Note that for the above, a student must show their understanding of the terms “categorical” and “quantitative,” by matching each example to one or other. Again, it is assumed the examples have not been previously seen.

Example 2.2: (Design) Taking a sample using a stratified sampling design is (circle all that apply):
(a) Generally more difficult than taking a simple random sample.
(b) Taking simple random samples within sub-groups of the target population.
(c) Systematically ignoring sub-groups of the population.
(d) Likely to remove problems with non-response.
(e) Best when sampling units within each stratum are similar.

In the above example, although students will have been told the underlying features of stratified sampling, it is assumed that the wordings above differ appreciably from those previously seen. Therefore, elements of comprehending aspects of the sampling design are required.

3.3 Apply

At level 3, students must be able to perform a procedure to which they have previously been exposed. In the context here, this typically means producing an answer such as a graphic, a statistic, or a P-value. The procedure involved may be routine, at least to an expert, though to the novice, skills at both levels 1 and 2 are required.

In this context, it is assumed that there is no reasonable ambiguity in the student’s mind as to which method is to be applied. We argue that where it is feasible to assume that the student needs to select amongst competing methods, at least some of the task falls at level 4 (analyse). This is congruent with the authors’ experiences: students may accurately apply a method if in no doubt as to which method to use, but may flounder if the question is not explicit. As Schau and Mattern (1997) remark, students say things like "I can use the t-test when I know I'm supposed to, but otherwise I don't have a clue." For instance, when presented with data in pairs, but when the question neither asks for a paired test nor includes the word “pairs,” some students will perform two-sample tests inappropriately. We judge the performance of the test to be at level 3, but the selection of the correct procedure to be level 4 (in cases when in the student’s mind a choice may exist).

Outlines of sample tasks at level 4 for the seven topics are provided in Table 4.

Table 4. Sample tasks for level 3

EDA	Predict outcomes based on a graph, create a specified graph, apply a formula.
Design	Find the number of levels of a factor in an experiment, conduct a study given the design.
Probability	Identify the sample space in a probability word problem, compute the variance of a random variable.
Sampling distributions	Use the CLT to approximate probabilities relating to a mean or sum.
Inferential methods	Use a specified method on given data, to perform a test or create a confidence interval.
Regression	Compute the regression line for given data, find a predicted value.
ANOVA	Compute or complete ANOVA given data or key summary statistics, find the P-value for the test.

Certain tasks may appear rather trivial to be assigned at this level, but one recalls that the taxonomy measures depth of mastery and not directly the level of difficulty. For instance, computing the variance of a data set requires recall of the formula, an understanding of how to perform the computation, and the ability to complete the computation correctly. The same is true for finding the estimate of the unknown within-group variance in the context of ANOVA. Whether these tasks are deemed difficult or not, they do require a level of cognition in addition to those found at levels 1 and 2.

The sample tasks for "Design" include conducting a study given the study's design – not practical in an examination. It is, however, one of the few generic tasks that could be prescribed on this topic at this level in an introductory course.

Two example questions at level 3 follow:

> Example 3.1: (Probability) Consider two events A and B within a sample space. Suppose $P(A)=0.4$, $P(B)=0.6$ and $P(A \cup B)=0.7$. Find $P(A \cap B)$.
>
> Example 3.2: (Sampling distribution) A drinks company firmly believes that 10% of all soft drink consumers prefer its brand, Simpsons' Water. To verify this, it conducts a survey of 2500 consumers, and is interested in the number who claim to prefer Simpsons' Water to all other brands. The managing director of the firm may reconsider the claim of the company if no more than 220 of the sample state a preference for Simpsons' Water. Use the Central Limit Theorem to estimate the probability of this event assuming the claim is correct.

The first question requires probability rules which we assume the students have applied previously. The second question would require thinking at level 4, were it not given to use the CLT. Note both questions are devoid of requirements to draw conclusions in the context of a study.

3.4 Analyse

At this level, tasks require students to perform a deductive step of some sort. This may take the form of selecting which method is most appropriate for a given setting from a set of methods previously seen. In such cases the application of the method would be at level 3, but selecting the right method can, for instance, demand an appreciation of the structure of the data that is not obvious from the way they are presented. Other examples require students to extract key information from a case study, such as formulating hypotheses and drawing conclusions in the context. Some sample tasks are described in Table 5.

Table 5. Sample tasks for level 4

EDA	Decide which EDA method is suitable in a given situation, draw conclusions in context.
Design	Identify sources of possible confounding and bias in a study.
Probability	Interpret the parameters in a probability model, propose a known model for a context.
Sampling distributions	Identify features that influence a sampling distribution, explain a particular sampling distribution in context.
Inferential methods	Decide with method is appropriate, recognise structure in data, formulate hypotheses and draw conclusions in context.
Regression	Explain the influence of an outlier, consider transformations, interpret in context.
ANOVA	Recognise when ANOVA is (in)appropriate, draw conclusions in context.

In many example questions, such as 4.1 below, level 4 cognition combines with level 3 skills. The second example, which is not from an introductory course, illustrates a deductive reasoning task in the context of regression models. Students had not previously met a scenario of the type described. In should also be noted in Example 4.2 the notation in the question was as presented here and was familiar to the students, with Greek letters indicating parameters and lower case letters denoting point estimates.

> Example 4.1: (Inference): In a certain city, 25% of residents are European. Suppose 120 people are called for jury duty, and 24 of them are European. Does this indicate that Europeans are under-represented in the jury selection system? Carry out an appropriate hypothesis test at the 10% significance level. Define the parameter(s) relating to your test, and clearly state your null and alternative hypotheses in the given context.
>
> Example 4.2: (Regression): Consider data on patient outcomes (Y) for patients who are randomly allocated to one of three drug treatments. Let A be a binary indicator taking the value one for patients allocated to the first drug, and zero otherwise. Similarly, let B and C be binary indicators for allocation to the second and third drugs respectively. Someone who is unsure about statistical methods ends up getting output from two fitted regression models:

> Model 1: $Y=\alpha_0+\alpha_1 A+\alpha_2 B+\varepsilon$,
> Model 2: $Y=\beta_0+\beta_1 A+\beta_2 C+\varepsilon$.
> Which of the following is/are guaranteed to happen? (Check all that apply.)
> (a) $a_0=b_0$
> (b) $a_1=b_1$
> (c) $b_2-b_1=-a_1$
> (d) The residual sums of squares from both models are the same.

3.5a Create

The "Create" level is characterized by the necessity for a student to apply their knowledge to devise a method, model, or evaluation procedure that is new, at least to the student. Tasks at this level require high-order thinking, and would rarely be assessed on examinations at an introductory level. The requirement to produce a novel artefact prevents cognition at this level being assessed by multiple choice questions. Some sample tasks are provided in Table 6, in which the terms "new" and "novel" describe items that were previously unknown to the student.

Table 6. Sample tasks for level 5a.

EDA	Invent a new summary statistic for a given purpose.
Design	Design a study to investigate a hypothesis.
Probability	Formulate a novel probability model for a context.
Sampling distributions	Draw inference based on a novel sampling distribution.
Inferential methods	Formulate novel hypothesis test for a given scenario.
Regression	Use regression to propose a model based on multiple predictors
ANOVA	Invent a multiple comparison test.

Examination questions involving invention are rare at the introductory level, though the use of invention tasks in statistics instruction has advocates (Schwartz and Martin 2004, for example). Part of a question requiring students to design a study is provided below as an example of a test question at level 5.

> Example 5.1: (Design) The two STAT 101 instructors notice that some students fall asleep in class. They plan to conduct a study to assess whether having coffee right before class improves their students' attention span. They know that the lecture time (9am and 2pm) may affect attention span, but they are not interested in assessing the lecture time effect. The 9am section has 197 students, and the 2pm section has 340 students. Design an appropriate experiment to establish a causal effect of coffee on students' attention span. Remember to take lecture time into account in designing the experiment. Draw a diagram that outlines an experiment. Do NOT describe the study in words.

3.5b Synthesise/Evaluate

Tasks at level 5b may require students to collate and interpret information from a variety of sources without explicit guidance as to what value to attribute to each source. Alternatively, or perhaps in addition, tasks may request a critique, comparing and contrasting methodologies, for example. Sample tasks at this level appear in Table 7.

Table 7. Sample tasks for level 5b.

EDA	Compare measures of centrality, critique EDA approaches, collate and summarise EDA information from multiple sources.
Design	Compare and contrast different study designs, critique a given study.
Probability	Compare different probability models, identify limitations of a model.
Sampling distributions	Compare use of asymptotic theory to an exact approach.
Inferential methods	Compare different tests for a given scenario, critique arguments for Bayesian and frequentist methods.
Regression	Compare output from different regression analyses, critique software for regression.
ANOVA	Compare ANOVA with alternative approaches.

Examination questions at this level are rare in undergraduate statistics; indeed, for some topics it is difficult to construct questions requiring synthesis or evaluative skills. In the review of various examination papers across several different courses, the authors met only one question at this level, that being from an advanced inference course. The question in part asked the students to compare a frequentist approach to a Bayesian alternative, both of which the students had applied. That rather lengthy question is not repeated here, although an assignment task from a third-year course is provided in the Appendix as a sample task at level 5b.

4. The Blooming process: A case study

4.1 The course

Before we describe the process via which the authors went about Blooming the examination papers considered here, some information is provided about the course on which the questions were used. The questions considered were from an introductory course at the authors' institution, a large publicly-funded university in North America. The course is a somewhat generic introductory one-term class in statistics, with the majority of the students being in science programs. The course is offered each term, including the summer term. In the two winter terms, up to fall 2012, either two or three sections were offered each term, typically of around 140 students each. Subsequently, fewer but larger sections have been offered, with two sections of around 280 students in the first term, and one section of about 320 in the second. When multiple sections are offered in the same term, all students sit the same final examination. The summer section attracts fewer students, usually around 80.

The three authors have each taught the course in question since 2003, though various other instructors have also been involved. Only examination papers were considered where an author had involvement in the setting of the questions. In the following, the first winter term is denoted

F (for “fall”), the second W, and the summer term S. Years are appended, so that, for instance, F11 denotes the fall term in 2011.

A motivation for our work was to evaluate the impact of pedagogical innovations adopted in the course since 2007. These reforms include the introduction of comprehensive learning outcomes (2007), personalised response system use in each class (clickers, also 2007), a revision of the laboratory activities (2008), the introduction of on-line homeworks via the WeBWorK system (2012), and optional study skills workshops (2013). Throughout the process of transforming the course, efforts were made to shift assessment away from testing the ability to perform routine calculations and towards assessing the appreciation of important concepts. Appraising the effectiveness of these reforms as gauged by student learning proved difficult, partly since changes have been incremental, and partly since no baseline data exist for comparisons. That said, some examination papers were available from prior to the reforms, along with records of students’ overall scores. It was of interest to see whether the examinations have increased in difficulty, as measured by Bloom’s Index, and also to explore student performance. The prior perception of the authors was that examinations had become more cognitively demanding, yet students were performing at least as well.

Largely out of interest, the first author computed the Bloom’s Index for an examination on a second course in statistics he taught. This course, referred to as “the second course” in what follows, is a follow-up from an introductory course such as the one described above. This second course recruits around 120 students per year, mostly third and fourth year science students, approximately a third of whom are statistics specialists. The course is taught once per year, in the fall term.

4.2 Applying the method

As a starting point, the first author produced descriptors of question styles and example questions at each level for various topics covered on the course, similar to the content of section 3 and the Appendix here. These were distributed to the other authors, along with some references on the taxonomy and a short summary of the key ideas. Following meetings in which the taxonomy was discussed in the context of an introductory statistics course, it was agreed that the authors would initially each independently attempt to Bloom the exam from the second term in 2008. This examination paper was chosen as a training paper as it fell in the middle of the time period over which examinations were available, and was representative of the time at which transformations to the course were underway but not fully implemented.

After each author independently Bloomed the W08 examination using prototype versions of the descriptors and sample questions provided here, the parties met and discussed discrepancies in their ratings. There was a need to refine the guidelines, in particular clarifying the distinctions between tasks at levels 2 and 4. The desirability of splitting questions into components at different levels – typically 3 and 4 – became apparent.

Computing the Bloom’s Index requires knowledge of the points weighting on parts of long-answer questions, and it was determined even a multiple-choice question may require skills at both level 4 (for instance, select the appropriate test) and level 3 (compute the test statistic).

Assigning such a task entirely at the highest level seems artificial, although admittedly correctly computing the wrong statistic would lead to no marks for a student in a multiple choice setting. It is a more accurate reflection of the processes required to answer a statistics question to break down the task into sub-tasks where a clear distinction can be made. For instance, in a question requiring the student to select an appropriate test, perform the test, and report a conclusion in the context of the study, the performance of the test is at level 3, whereas the selection of the method and the conclusions are at level 4. This in part reflects how students are graded: for instance, a student performing a two-sample t-test where a paired t-test was appropriate could still obtain some credit if carrying out the (incorrect) test correctly. More importantly, there are cases where such subdivision of tasks better reflects the cognitive process required.

Subsequent to the refining of the guidelines, the authors independently Bloomed examinations from S03, W06, W11, W13, and F13. In each case one of the authors had been involved in teaching the course and setting the examination, and so could provide information about which questions, if any, the students had previously seen and therefore are at level 1. The W13 paper was Bloomed after the others, and following the introduction of two additional rules, the need for which had arisen during the Blooming process and had resulted in discrepancies between at least two raters. These are described below.

The W11 examination included (as Q5) the following:

> Q5. Here are three data sets:
> Data I: 5, 7, 9, 11, 13, 15, 17
> Data II: 5, 6, 7, 11, 15, 16, 17
> Data III: 5, 5, 5, 11, 17, 17, 17
> Which of the following statements is/are true about the three data sets?
> (a) The three data sets have the same range.
> (b) The three data sets have the same standard deviation.
> (c) Data III has the largest standard deviation among the three data sets.
> (d) Both (a) and (b).
> (e) Both (a) and (c).

An expert would not compute the relevant standard deviations to answer this question, instead applying an analytical process to assess which data set had the larger statistic. No doubt many students simply computed the statistics and so needed only level 3 skills, leading to the decision that by default the level of the least complex approach possible should be attributed to a question where alternative approaches are feasible. The suggestion is that if an approach to a question can successfully and practically be performed at a low level of mastery (as per Bloom's) then many students would adopt that rather than consider more expert-like approaches that may be simpler in some way, and the task should be assigned to the lower level. This guideline appears sensible independent of the level of the course.

The second issue that arose was when subsequent parts of a question provide information that reduces the level of earlier parts. For instance, in the F13 exam, questions 32-35 related to the same case study. Only Q35 revealed that ANOVA was an appropriate method to apply. Presumably, though, students that had the wrong idea would revisit their answers, enlightened by the information in Q35, so that level 4 cognition was not required for Qs 32-34. However, this does assume a student can recognize that new information is relevant to previous responses, which requires some level of mastery. In that sense, any change in level in responding to an earlier question relies on the nature of the information appearing in a subsequent question. Such a change would not apply in testing situations that prevent students returning to earlier questions, but otherwise it is a nuisance when attempting to treat related questions as stand-alone entities. The authors believe scenarios similar to the one described in the F13 exam would be the most common circumstance where this problem would arise.

Initially, complete agreement between the three raters on the W08 examination occurred on 60% of items. Once issues such as those mentioned had been resolved, concordancy between the three raters was relatively high across the exams considered. All three raters agreed on around 70% of items, and all three disagreed on less than 5%. These levels of agreement compare favourably with those of Fairbrother (1975), Crowe et al. (2008), Freeman et al. (2011), and Casagrand and Semsar (2013).

Apart from cases where disagreements could be resolved via introducing a new rule on which all agreed, for the purposes of computing Bloom’s Indices, the rules of Zheng et al. (2008) were applied: when two raters agreed their majority choice was adopted, and in cases where the three raters disagreed sequentially (say, assigning 2-3-4 ratings), the middle level was adopted. There were no cases of non-sequential discrepancies. Overall, the descriptors and sample questions were deemed by the authors to be adequate to consistently provide a higher level of concurrency than is reported in comparable research, and the construction of a dichotomous key was considered neither viable nor necessary.

No guide or rubric for Blooming can be definitive, and certain types of questions remained problematic to place on the taxonomy. Questions on EDA, perhaps surprisingly, were amongst those where the authors disagreed most. Maybe on this topic it is hard not to conflate depth of cognition with level of difficulty. To give an example, one W13 question asked students to provide an estimate of the sample mean of a data set based on a histogram. Formulae do exist for this task, but none had been shown to the students and it was not expected that they would concoct their own formula, which would be a level 5 (Create) task. Nonetheless, such an approach would be viable. More reasonably, students could be expected to apply their understanding of the histogram to arrive at a plausible estimate, yet this “apply” task is very different from a relatively straightforward application of a formula, and arguably requires analytical thinking. Such an example illustrates that different approaches to a question could require very different levels of thinking, and even categorising all “apply’’ tasks within an introductory course is not easy.

4.3 Results

Table 8 below presents results from the six examinations considered for the introductory course. A noticeable feature that changed across time is the proportion of total marks assigned to multiple choice questions, and relevant data are provided. In Table 7, %MC is the percentage of total grade from multiple choice questions, whereas %LOCS is the percentage of distinguishable question parts assessed as requiring LOCS, with %HOCS the corresponding percentage for HOCS.

Table 8. Data on six introductory statistics examinations, including summaries of student performances.

				Exam			
		S03	W06	W08	W11	W13	F13
Exam	%LOCS	36.7%	36.6%	32.7%	37.1%	29.0%	18.2%
Data	%HOCS	63.3%	63.4%	67.3%	62.9%	71.0%	81.8%
	%MC	21%	16%	40%	70%	65%	85%
	BI	59.6	56.2	61.5	58.6	64.4	67.4
Student	N	69	168	127	157	323	289
data	Mean%	77.1%	69.8%	63.4%	71.4%	67.8%	64.2%
	Std. dev.	12.5%	14.4%	13.1%	13.1%	15.4%	15.1%

Data were included for sections of the course where one of the authors was an instructor. In computing summary statistics for student performances, only students sitting the final examination were included. Since summary statistics are based on all students available, no inferential methods are provided here. Only in the rather dubious sense of the students in each section being a subset of all possible students who may have chosen that section can the observations be considered a sample.

The single examination paper from the second course had a Bloom's Index assessed at 60.8, with 20% of the parts only requiring LOCS. Multiple choice questions comprised 32% of the marks available. In total, 129 students sat the examination, with a mean grade of 60.8% and a standard deviation of 9%.

4.4 Conclusions

Taken as a whole, the reported statistics for the examinations considered compare favourably with similar findings by other authors. Freeman et al. (2011) provide figures for four test papers on each of six runs of a biology class taken mostly by second year students at University of Washington. The Bloom's Indices are standardised (by multiplication by 6/5) to compare with our figures (for the course labelled "Intro Stat") in Table 9. Final examinations considered by Freeman et al. were assigned BI values between 54.1 and 66.4, slightly lower than the range in Table 8.

Table 9. Comparison of results with Freeman *et al.* (2011).

Course		Bloom's Index	
		Minimum	Maximum
	Intro Stat	56.2	67.4
	Biol Midterm	43.3	70.7
	Biol Final	54.5	66.4

Zheng et al. (2008) used three raters to compare 109 biology questions from the North American Medical College Admission Test (MCAT) with similar numbers of questions from other sources, including AP Biology, a sample of introductory biology test questions from three universities, the biology Graduate Record Examination (GRE), and five first-year medical school courses from a single institution. The percentages of marks on HOCS for each was taken as a proportion of the marks available from the questions sampled from each type of exam. Results are given in Table 10, below. These figures reported in Zheng et al. (2008) are low compared to the examinations considered here, where, for instance, the W13 and F13 papers have respectively 77% and 82% of total marks on HOCS.

Table 10. Comparison of results with Zheng *et al.* (2008).

		Mean %HOCS
Course	Intro Stat	68%
	MCAT	45%
	AP Biology	36%
	Undergrad Biol	51%
	Medical School Biol	21%
	GRE	35%

If one regresses mean exam score against Bloom's Index for the introductory statistics exams considered, one cannot reject the hypothesis that the slope is zero, contradicting the expected result that mean score would decrease as BI increases. Assuming our Blooming methods have been objective and consistent, this lack of inverse relationship between performance and difficulty does provide some (admittedly weak) support for the notion that changes to the pedagogy in the course have been effective.

Results in Table 8 do indicate that examinations on the course increased in difficulty level over time, as measured by both BI and the percentage of questions requiring high order cognitive skills. In particular, the two most recent examinations considered had the highest scores on both these measures. Since eleven WeBWorK on-line homeworks were introduced in 2012, it may be that, consciously or otherwise, instructors are setting more taxing examinations in the knowledge that most students have experienced much in the way of regular, relevant, formative assessment during the course. That regular, spaced assessments are beneficial to learning compared to fewer, high-stakes tests, is widely accepted. See, for instance, Myers and Myers (2007) for a study involving a statistics course. The possible impact of other interventions is harder to discern, if indeed it is sensible to isolate individual effects.

A feature that was not initially considered but became apparent on inspecting the examinations was the great variation in the percentage of marks available from multiple choice questions. This percentage rose dramatically over the period of interest, from as low as 16% in W06 to a high of 85% in F13. There are reasons for this change, and possible consequences. As student numbers have increased without a commensurate increase in teaching assistant support, the need for a speedy turnover of examination grading leads, pragmatically, to an increased use of multiple choice questions that can be machine graded. Yet, there are pedagogical reasons why multiple choice questions are more favoured now than previously. In recent years, the authors have become aware both of many student misconceptions and the ability of multiple choice questions to explicitly tease out misunderstandings in a way that can be more difficult with free-form response questions. One author, for instance, had earlier held the opinion that multiple choice questions were able to assess only low order skills, but now appreciates that cognitive processes at levels 3 and 4 can often be accurately assessed with good multiple choice questions. Concept inventories in statistics and other disciplines, such as the CAOS test (delMas et al. 2007), are comprised entirely of multiple choice questions.

Increasing the proportion of marks derived from multiple choice questions may, we conjecture, have the by-product of making a test harder for students. On a “long answer” question, a student with faulty reasoning may nevertheless pick up partial credit. For instance, at the authors’ institution, on a long-answer question requiring a paired t-test, a student who erroneously performed a two-sample t-test could nevertheless obtain some partial credit if completing his or her test correctly. However, a student applying faulty reasoning on a multiple choice question will gain no marks, since the flawed thinking would almost certainly result in the selection of a distractor.

One might argue that students can gain marks on multiple choice questions by merely guessing, which is true in the sense that were students to purely guess they could be expected to obtain around 20% of the total marks. In our experience, however, students rarely appear to guess, since they enter the examination with at least partial knowledge of the topics assessed. In this way multiple choice questions may in practice make gaining marks more difficult for students with partial understanding compared to free form-response questions assessing cognitive skills at the same level. The authors are not aware of any research that has explored this conjecture, though there is on-going work on awarding partial marks for multiple choice questions (Day, personal communication, June 9, 2015). If the hypothesis can be entertained, however, the examinations considered here have increased in difficulty by %MC, a measure additional to BI and %HOCS.

The BI does not distinguish between multiple choice questions and other styles of questions, except where it is impossible to test an outcome at a higher level by multiple choice. The "objective" level of difficulty of a test (as measured by the BI, for example) should not be confused with measures based on student performance. If our conjecture above is correct - that a higher proportion of marks for multiple choice questions can make a statistics test harder for students with respect to gaining marks for the same level of mastery - then the original Table 8 certainly shows an increase in difficulty over time due to the rise in the proportion of marks from multiple choice questions. Whether the conjecture is correct is unclear, and our data could not shed light on the matter as teaching methods and resources changed over time.

Student performance has been quite variable over the years, but contrary to what one might expect, there is no significant relationship between mean grade and Bloom’s Index. A somewhat anomalous result appears from summer 2003, the earliest class considered, where the mean grade of 77.1% is the highest observed. Students taking courses over the summer session rarely take a full course load, typically sitting just one or two courses during that term. Hence such students may be able to devote more time and energy to each course taken during that period. Both anecdotal evidence within our department and more detailed analysis by colleagues in other departments at our institution support the hypothesis that summer students on average perform better on a course than students taking the same course during a winter term. There may be demographic differences with summer students too, and feasibly the smaller class size has some impact. Hence, perhaps a more reliable baseline is the W06 cohort.

The negative correlation between %HOCS and mean score (-0.754) on all exams considered (including the second course) does suggest that students find exams more difficult with a higher proportion of marks awarded for higher order skills. The corresponding regression is just significant at the 5% level, giving further support to the hypothesis. This concurs with the broad conclusions of other researchers, arguably adding finer detail.

The data on the second course are of interest as they appear to dispel the suggestion that examinations on high level courses must make more cognitive demands on students than those on lower level courses. The notion is true in the sense that students need to master at least most of the concepts in the first course to pass the second, so by that measure at least the second course is harder. However, the final examination did not appear more demanding as measured by BI than recent tests on the introductory course. The %HOCS on the second course examination is relatively high, however, and further work may indicate that this is a feature in upper level courses compared to those at lower levels.

5. Discussion

We have presented suggestions as to how to interpret Bloom's taxonomy in the context of statistical education. Although previous researchers have referred to Bloom's as a means of calibrating the levels of mastery required to perform task in statistics, there does not appear to have been an attempt to make explicit how to align assessment tasks on the taxonomy's scale. With a particular focus on the type of tasks learners encounter in introductory courses, we have suggested guidelines for allocating test questions on Bloom's taxonomy. The post hoc evaluation of student performances on a course (or indeed across courses) is proposed as a possible application of this Blooming approach. As an example of this application, a case study is described involving an introductory course at the authors' institution, where certain practical implications and potential difficulties with the Blooming method were encountered and discussed.

Although the authors believe the guidelines presented for assessing the Bloom's level of statistical tasks are practical to apply, there are inevitably caveats. It is impossible to discern the extent to which students have prepared for particular tasks, for instance, even if full knowledge is available as to what materials were provided by the instructor. In coining the term "push-down effect," Merrill (1971, p.38) pointed out "Learners have an innate tendency to reduce the cognitive load as much as possible; consequently a learner will attempt to perform a given response at the lowest possible level." The authors find that many students hanker for more worked examples in our courses, and conjecture whether the motivation is implicitly that repeated exposure to completed questions enables learners to "push down" the cognitive load when attempting similar tasks later. Experts, by virtue of increased experience, are able to "push down" their cognitive load when performing tasks that to a novice would be highly demanding. Hence it can be impossible to arrive at entirely objective assessments of the Bloom's level of certain undergraduate examination questions.

Caveats aside, the approach described here has potential for various uses. As presented, Blooming examinations can help instructors compose examinations to test concepts across levels of mastery, and assess trends in difficulty levels in their examinations both within and between courses. Blooming final examinations can provide a relatively easy method for assessing the impact on student attainment due to pedagogical interventions. A difficulty in such research can be obtaining suitable baseline data, since it is often only subsequent to teaching a course by one method that an instructor decides to implement a new approach, by which time the students who had been taught by the previous method have completed the course and moved on. Assuming past test papers and student performance data are available, however, our approach can lead to a means by which course modifications can be assessed in such cases.

In addition to assessing relative difficulties of examinations, the Bloom level descriptors may be useful as a teaching aid. Crowe et al. (2008) illustrate the use of Blooming in enhancing the teaching and learning in biology classes, in part by making students aware that LOCS will not be sufficient for them to succeed in the class. This strategy appears particularly applicable in statistics teaching where, in the authors' experience, some students perceive success in the discipline to be overly reliant on the ability to recall and apply rules and formulae. Indeed, activities that ask students to Bloom particular problems may have benefits in statistical education, following the work of Bissell and Lemons (2006) in biology classes.

Alerting students to the depths of cognitive processes required in performing statistical tasks may also have the side effect of altering student behaviour when attempting to learn the subject. For instance, Scouller (1998) found that study habits of education students differed depending on the perceived cognitive demands of the assessment tasks. The benefits of appreciating levels of mastery in statistics may only occur to students if presented within a statistics course if, as some educators such as Wingate (2006) argue, efforts to improve students' study skills *per se* are useless if detached from the discipline in question.

It is hoped that the attempts here to describe and pseudo-objectively categorize the depth of thought processes required to solve problems in the statistical sciences may assist teachers in refining their curricula and assessment tools and help learners in developing their meta-cognitive skills within the discipline.

Appendix

Sample questions are provided for each of the six levels of Bloom's taxonomy. Most of the questions are actual test questions that have been used at the authors' institution. Comments are provided where required.

A. Knowledge/Remembering

It is assumed for questions 1, 2, 5, and 6 that students have been told each of the facts that are correct, and in not recalling the other statements determine those false. It might be argued, however, that level 2 thinking is being applied in this process. This is most likely in the first question.

1. (Design): Indicate which of the following is/are true for clinical trials that are *double-blind*, by circling the corresponding letter(s):
(a) The placebo effect is eliminated.
(b) Neither subject nor medical staff knows which treatment is applied.
(c) They are most appropriate for matched pair designs.
(d) The data will be analysed without regard to which treatments were applied.
(e) The results are encoded to "blind" information about the subjects.

2. (Probability): Indicate by circling the appropriate letter(s) which of the following statements is/are always true for a Normal distribution:
(a) It is a symmetrical distribution.
(b) Its mean and standard deviation are similar.
(c) It is a skewed distribution.
(d) It is good for modelling categorical variables.
(e) It is unimodal.
(f) Its mean and median are identical.

3. (Sampling distribution): State, but do not prove, the Central Limit Theorem.

4. (Inferential methods): Provide the formula for the test statistic for a one-sample t-test, defining all notation you use.

5. (Regression): The slope of a regression line and the correlation are similar in the sense that (choose all that apply):
(a) they both have the same sign.
(b) they do not depend on the units of measurement of the data.
(c) they both fall between -1 and 1 inclusive.
(d) neither of them can be affected by outliers.
(e) both can be used for prediction.

6. (ANOVA): In performing the usual hypothesis test in the analysis of variance using the mean-square ratio, the alternative hypothesis is that
(a) at least two of the underlying group means are different.

(b) all the underlying group means are different.
(c) all the within-group variances are equal.
(d) at least two within-group variances are different.
(e) all the within-group variances are different.

B. Comprehend/Understand

1. (Design): Taking a sample using a stratified sampling design is (circle the corresponding letter(s) of all that apply):
(a) Generally more difficult than taking a simple random sample.
(b) Taking simple random samples within sub-groups of the target population.
(c) Systematically ignoring sub-groups of the population.
(d) Likely to remove problems with non-response.
(e) Best when sampling units within each stratum are similar.

2. (Probability): Ignoring twins and other multiple births, suppose that the babies born in a hospital are independent, with equal probabilities that a baby is born a boy or girl. Consider the events A={the next two babies born are both boys} and B={at least one of the next two babies born is a boy}. Are A and B independent?

3. (Sampling distribution) A survey of people who own stocks investigated how frequently the respondents monitored the value of particular stocks. Of the 545 respondents, 386 indicated they checked the values of particular stocks at least once a day. What is a parameter of interest in this study?

4. (Inferential methods): A scientist is interested in investigating a physical parameter θ which can take one of two possible values. She plans to conduct a test of $H_0:\theta=\theta_0$ against an alternative hypothesis $H_a:\theta=\theta_a$. She constructs her test at the 1% significance level and to have power 0.90. What is the probability she commits a type I error?

5. (Regression): In a regression model, does the standard error of the estimate of the slope depend on the values of the response variable?

6. (ANOVA): In a comparison of gas mileage per gallon, measurements were taken on 10 Honda Civics, 15 Toyota Yaris's and 30 Mazda 3's. Name one of the parameters of interest.

C. Apply

1. (EDA): The following give the expected number of miles per gallon for ten vehicles, in highway conditions: 27, 28, 24, 28, 30, 22, 30, 28, 31, 38. Find the mean, median, and mode of the values.

2. (Inferential methods): In a two-sided significance test for a mean, the test statistic was -2.12 which is expected to be a value from the standard Normal distribution under the null hypothesis. Find the p-value of the statistic.

3. (Regression): A mining company took twenty samples of sediment from the ocean floor. In each sample, the quantity of Uranium, Y, and the amount of the mineral feldspar, X, were recorded. The mean amount of feldspar was 10.14 micrograms (μg) and the mean amount of Uranium was 9.86 μg. The variances of the variables were 198.85 and 95.90 respectively (in μg^2). The sample correlation between the two variables was 0.72. Find the regression line of Y on X here.

4.(ANOVA): A study investigated whether month of birth impacts on the time a baby learns to crawl. Parents with children born in January, May or October were asked the age, in weeks, at which their child could crawl one metre within a minute. The data are summarised below:

		Crawling age		
		Mean	St. dev.	Size
Birth Month	January	29.84	7.08	34
	May	28.58	8.06	29
	October	33.83	6.93	40

The data from each birth month are assumed to follow a Normal distribution. The analysis is via ANOVA, with an incomplete ANOVA table given below:

Source	Sums of squares	DoF	Mean Square	F
Between groups	505.26			
Error			53.45	
Total				

Compute the test statistic for the test.

D. Analyse

In the third question below, only the selection of the model (Binomial here) is at level 4, the remainder of the task is at level 3. For the fourth question, it is assumed that the students have not been told about the asymptotic properties of the sample variance, and must deduce that the CLT is applicable. As an aside, the authors find that even after being informed about the convergence of the sample variance, most students answer this question incorrectly.

Question 6 is the first four parts of a multi-part question from the second course. The first two parts are at level 3.

1. (EDA) Twenty specimens of a type of polymer were used in an experiment. Ten were randomly assigned to be exposed to an accelerated aging process, the remainder comprising a control group. Measurements of the tensile strength were made on each specimen, the data below being in psi.

Control group:	227, 222, 218, 217, 225, 218, 216, 229, 228, 221 221
Accelerated aging:	219, 214, 215, 211, 209, 218, 203, 204, 201, 205 205

Create a suitable graphic to compare the three samples. Would you say the accelerated aging process has an impact on the tensile strength of the polymer? Justify your answer carefully.

2. (Design): A study investigated a possible relationship between TV viewing and school grades amongst a sample of 600 high school students. The study indicated that there is a correlation of -0.85 between the number of hours spent watching TV per week and the mean grade score. Circle which of the following is/are likely to be true, and explain your answer:
(a) TV viewing causes poor school performance.
(b) The sample size was too small to provide any reliable evidence.
(c) There are variables which are confounded with TV viewing habits.
(d) The correlation here does not indicate any possible relationship.
(e) Watching more TV would increase a student's high school grades.

3. (Probability): The probability that a driver of a certain age will pass the driving test at the first attempt is 0.35. If six such drivers are entered for their first tests on a given day, find the probability that at least two pass.

4. (Sampling distribution): The variance of a sample is usually defined as $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. You then take a random sample of size 150 from the same population, and again compute the sample variance.
Which of the following best describes how we should expect the two sample variances to compare?
(a) The two variances would be identical.
(b) The second variance would be expected to be larger than the first.
(c) The second variance would be expected to be smaller than the first.
(d) We would not expect the second variance to be much bigger or much smaller than the first.
(e) It is impossible to say anything about the relative values of the two variances.

5.(Inferential methods): So-called racial steering occurs when a real estate agent has a tendency to show prospective property buyers only homes in neighbourhoods already dominated by the race of the prospective buyer. In a court case alleging that a real estate agent was applying racial steering to renters, data were provided about how the race of a potential renter of an apartment in a complex appeared to influence the section in the complex the realtor showed the potential renter. The complex split into two sections, section A which was predominantly white and B which was predominantly black. The data for a two-month period are below:

		Potential Renter Race	
		White	Black
Section shown	A	9	1
	B	8	3

(a) What hypothesis (or hypotheses) is/are of interest to test in this context? (Choose all that apply.):

i. There was no difference between the number of white and black potential renters.

ii. There is no dependence between the race of a potential renter and the section they are shown by the realtor in the complex.

iii. White potential renters are just as likely to rent an apartment in the complex as black potential renters.

iv. The chance of being shown an apartment in each section does not depend on the race of the potential renter.

v. There was no difference in the numbers of apartments being available to rent in the two sections of the complex.

(b) Select a hypothesis you chose above. Explain why in this context it would be reasonable to adopt a one-sided alternative hypothesis.

(c) In conducting a hypothesis test here, we might apply a Chi-squared test. What reservation(s) would you have about applying such a test to the data given in the table? (Choose all that apply.)

i. At least one of the expected cell counts is very small under the null hypothesis.

ii. The test is not suitable for testing an interesting hypothesis in the context of data such as those in the table.

iii. It is known that when the counts in the table are small the test tends to have an actual significance level much larger than the level selected, so that the null hypothesis is rejected too often when it holds.

iv. It is known that when the counts in the table are small the test tends to have an actual significance level much smaller than the level selected, so that the null hypothesis is rejected too rarely when it holds.

v.It is known that when the counts in the table are small the test tends to have an actual significance level much smaller than the level selected, so that the null hypothesis is rejected too often when it holds.

6. (Regression): In an experiment on study habits and the relation to final exam grades, data were collected on the scores on the final examination and the estimated hours spent revising for each of the forty students on a course. Some of the students reported that most of the time they spent revising was in the presence of some form of distraction, such as a TV or radio. The remaining students studied most of the time with no such distractions. It is of interest to model how the final test score, Y, depends on the amount of hours spent revising (X say). A model of the form

$$Y=\beta+\beta_1X+\beta_2Z+\beta_3XZ+\varepsilon$$

was fitted, where ε is a Normally distributed error and the variable Z is 0 if a student was not mostly revising in the presence of a distraction and 1if a student was mostly revising in the presence of a distraction.

The following estimates and standard errors were obtained:

Parameter	Estimate	Standard error
β_0	29.101	2.879
β_1	0.885	0.221
β_2	-0.0113	0.0350
β_3	-0.389	0.0222

(a) Based on the model fitted above, for a student who revised mostly without distraction, by how much would you predict their grade to increase for each additional hour of studying?

(b) Based on the model fitted above, for a student who revised mostly with a distraction, by how much would you predict their grade to increase for each additional hour of studying?

(c) For each of your answers in (a) and (b), either provide an approximate 95% confidence interval for your estimate, or explain why you are unable to do so.

(d) For the model fitted above, it is of interest to test whether β_2=0. State clearly in words what this hypothesis is testing in the context given.

7. (ANOVA): A pharmaceutical company conducted a clinical trial to investigate the effectiveness of three drugs in treating HIV. Three hundred HIV patients were randomly selected from the Immunodeficiency Clinic at a local hospital. These patients were randomized to receive one of the three drugs (Combivir, Stribild, and Tivicay). To monitor how the patients were responding to the drugs over a two-year period, they had their viral load (a quantitative measure of the severity of HIV infection) measured at start of the trial and every six months thereafter. A lower viral load indicates a better response to the drug.

In order to compare the true mean viral load among the three drug groups of patients at the end of the two-year period, which of the following hypothesis tests is the most appropriate?

(a) A one-proportion z-test
(b) A paired t-test
(c) A two-sample t-test
(d) An ANOVA F-test
(e) None of the above is appropriate.

E. Create/Synthesise/Evaluate

Examination questions at level 5a and 5b are sparse at the undergraduate level, and rare in introductory courses. The question below is part of an assignment question, the data being from Roberson et al. (1997). The final part asks for an evaluation of the methods applied, and addresses some subtle points in the comparison of the tests performed.

> (Inferential methods): Students complete an on-line attitudinal survey both before and after taking an introductory Statistics course. One of the items on the survey asks students to rate their agreement with the statement "I expect to perform well on Statistics courses," responses being on a Likert scale, 1 indicating "Strongly disagree" and 5 indicating "Strongly agree." Suppose in a class of seventy students the responses before and after were as follows:
>
> | | **Responses** | | | | |
> |---|---|---|---|---|---|
> | | 1 | 2 | 3 | 4 | 5 |
> | **Before** | 0 | 67 | 3 | 0 | 0 |
> | **After** | 51 | 0 | 0 | 3 | 16 |
>
> In the above data, 51 students moved from 2 to 1, 3 moved from 3 to 5, 13 moved from 2 to 5, and 3 moved from 2 to 4.
> (a) Describe briefly and clearly what you consider to be the motivation for gathering the survey data described above.
> (b) Apply a t-test to the data above. Describe your test clearly, indicating the test used, the null and alternative hypotheses, the test statistic, the P-value, and the conclusion.
> (c) Strictly speaking, what assumptions does the above test require?
> (d) Do the assumptions of your t test appear plausible here? Explain your thinking clearly.
> (e) Perform a sign test on the data provided. Describe your test clearly, indicating the hypotheses, the test statistic, the P-value, and the conclusion.
> (f) Explain how to perform Wilcoxon's signed rank test on these data. Compute the test statistic and P-value for this test for the data given.
> (g) Comment on the suitability of the three tests you have performed. Decide which test you would prefer, clearly explaining your reasoning.

Acknowledgements

This work was funded in part by the Carl Wieman Science Education Initiative (CWSEI). The authors wish to thank all those involved with CWSEI for their support and feedback during this study, in particular Warren Code, Sarah Gilbert, and Ed Knorr. In addition, the authors are grateful to the two referees and the editor who provided valuable feedback on an earlier draft of this document.

References

Alacaci, C. (2004), “Inferential Statistics: Understanding Expert Knowledge and its Implications for Statistics Education.” *Journal of Statistics Education* Vol. 12 No.1. Available at http://www.amstat.org/publications/jse/v12n2/alacaci.html.

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., and Witmer, J. (2012), GAISE College Report, American Statistical Association. Available at http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf.

Allen, D., and Tanner, K. (2002), “Approaches to cell biology teaching: questions about questions.” *Cell Biol. Edu.* Vol.1, 63-67.

Allen, D., and Tanner, K. (2007), “Putting the horse back in front of the cart: using visions and decisions about high-quality learning experiences to drive course design.” *CBE Life Sci. Edu.* Vol.6, 85-89.

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., and Wittrock, M.C. (2001), *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). New York: Longman.

Bailin, S. (2002), “Critical thinking and science education.” *Sci. Ed.* Vol. 11, 361-375.

Begg, A. (1997), “Some emerging influences underpinning assessment in statistics.” In *The Assessment Challenge in Statistics Education*. Gal, I. and Garfield, J.B. (edit.) IOS Press, 17-25.

Bissell, A.N., and Lemons, P.P. (2006), “A new method for assessing critical thinking in the classroom.” *BioScience* Vol. 56, 66-72.

Bloom B. S., Engelhart, M.D., Furst, E.J., Hill, W.H., and Krathwohl, D.R (1956), *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain.* New York: David McKay Co Inc.

Casagrand, J., and Semsar, K. (2013), “Don't have a pre-post test, use a Bloom's rubric! The development and validation of a rubric for "blooming" assessments to measure student learning.” Available at http://www.colorado.edu/sei/documents/publications/IPHY/Blooms-rubric.pdf.

Chance, B.L. (2002), “Components of Statistical Thinking and Implications for Instruction and Assessment.” *Journal of Statistics Education* Vol. 10 No.3. Available at http://www.amstat.org/publications/jse/v10n3/chance.html.

Crowe A., Dirks, C., and Wenderoth, M.P. (2008), “Biology in bloom: Implementing Bloom’s taxonomy to enhance student learning in biology.” *CBE Life Sci Edu.* Vol. 7, 368-381.

delMas, R. , Garfield, J., Ooms, A., and Chance, B. (2007), "Assessing Students' Conceptual Understanding After a First Course in Statistics." *Statistics Education Research Journal*, Vol. 6, No. 2, 28-58. Available at http://iase-web.org/documents/SERJ/SERJ6%282%29_delMas.pdf.

Ebel, R.L. (1965), *Measuring Educational Achievement.* Prentice Hall, N.J.

Fairbrother, R. W. (1975), "The reliability of teachers' judgment of the abilities being tested by multiple choice items." *Educational Research*, Vol. 17, No. 3, 202-210.

Freeman, S., Haak, D., and Wenderoth, M.P. (2011), "Increased course structure improves performance in introductory biology." *CBE—Life Sci. Edu.* Vol. 10, 175–186.

Garfield, J.B. (1994), "Beyond Testing and Grading: Using Assessment To Improve Student Learning." *Journal of Statistics Education* Vol.2 No.1. Available at http://www.amstat.org/publications/jse/v2n1/garfield.html.

Garfield, J.B. and Chance, B. (2000), "Assessment in statistics education: Issues and challenges." *Mathematical Thinking and Learning*, Vol. 2, 99-125.

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G.W., Chance, B.L., and Holcomb, J.P. (2011), "Rethinking Assessment of Student Learning in Statistics Courses." *The American Statistician,* Vol. 65, No. 1, 1-10.

Hubbard, R. (1997), "Assessment and the Process of Learning Statistics." *Journal of Statistics Education* Vol. 5 No.1. Available at http://www.amstat.org/publications/jse/v5n1/hubbard.html.

Lane-Getaz, S. J. (2013), "Development of a Reliable Measure of Students' Inferential Reasoning Ability." *Statistics Education Research Journal* Vol. 12, No.1. Available at http://iase-web.org/documents/SERJ/SERJ12%281%29_LaneGetaz.pdf.

Lemons, P.P., and Lemons, J.D. (2013), "Questions for assessing higher-order cognitive skills: It's not just Bloom's." *CBE Life Sci Educ*. Vol.12, 47-58.

Knecht, K.T. (2001), "Assessing cognitive skills of pharmacy students in a biomedical sciences module using a classification of multiple choice item categories according to Bloom's taxonomy." *Am. J. Pharm. Ed*. Vol.65, 324–334**.**

Kropp, R.P., and Stoker, H.W. (1966), *The construction and validation of tests of the cognitive processes described in the "Taxonomy of educational objectives."* (Cooperative Research Project N.2117) US Office of Education.

Madaus, G. F., Woods, E. M., and Nuttall, R. L. (1973), "A causal model analysis of Bloom's taxonomy." *American Educational Research Journal*, Vol. 10, 253-262.

Marriott, J., Davies, N. and Gibson, L. (2009), "Teaching, Learning and Assessing Statistical Problem Solving." *Journal of Statistics Education* Vol.17 No.1. Available at http://www.amstat.org/publications/jse/v17n1/marriott.html.

Marzano, R.J. (2001), *Designing a New Taxonomy of Educational Objectives.* Corwin Press, Thousand Oaks, CA.

Merrill, M.D. (1971), "Necessary psychological conditions for defining instructional outcomes." *Educational Technology,* Vol.11, No.8, 34-39.

Myers, C.B., and Myers, S.M. (2007), "Assessing Assessment: The effects of two exam formats on course achievement and evaluation." *Innov. High. Educ.* 31, 227-236.

Prince, M. (2004), "Does Active Learning Work? A Review of the Research." *Journal of Engineering Education* Vol. 93 No.3, 223-231.

Roberson, P. K., Shema, S.J., Mundfrom, D.J., and Holmes, T.M. (1997), "Analysis of paired Likert data: How to evaluate change and preference questions." *Family Medicine* 27 No. 10, 671-675.

Schau, C. and Mattern, N. (1997), "Assessing Students' Connected Understanding of Statistical Relationships." In *The Assessment Challenge in Statistics Education*. Gal, I. and Garfield, J.B. (edit), IOS Press. 91-104.

Schwartz, D.L., and Martin, T. (2004), "Inventing to prepare for future learning: The hidden efficiency of encouraging student production in statistics instruction." *Cognition and Instruction* Vol.22, No.2, 129–184.

Scouller, K. (1998), "The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay." *Higher Education* Vol. 35, 453-472.

Seddon, G. M. (1978), "The properties of Bloom's taxonomy of educational objectives for the cognitive domain." *Review of Educational Research*, Vol. 48, No.2, 303-323.

Stanley, J. C., and Bolton, D. T. (1957), "A review of Bloom's "Taxonomy of educational objectives" and J. R. Gerberick's "Specimen objective test items, a guide to achievement test construction."" *Educational and Psychological Measurement*, Vol. 17, 631-634.

Steinhorst, R. K., and Keeler, C. M. (1995), "Developing Material for Introductory Statistics Courses from a Conceptual, Active Learning Viewpoint." *Journal of Statistics Education*, Vol. 3(3). Available at http://www.amstat.org/publications/jse/v3n3/steinhorst.html.

Tallman, M., and Carlson, M. P. (2012), "A characterization of calculus I final exams in U.S. colleges and universities." In *Proceedings of the 15th Annual Conference on Research in Undergraduate Mathematics Education* Vol. 2, pp. 217-226, Portland, Oregon.

Wingate, U. (2006), “Doing away with ‘study skills’.” *Teaching in Higher Education*, Vol. 11, Issue 4, 457-469

Zieffler, A., Garfield, J., Alt, S. Dupuis, D., Holleque, K, and Chang, B. (2008), “What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature.” *Journal of Statistics Education.* Vol. 16 No.2. Available at www.amstat.org/publications/jse/v16n2/zieffler.html.

Zheng, A.Y., and Lawhorn, J.K., Lumley T., Freeman, S. (2008), “Application of Bloom’s taxonomy debunks ‘the MCAT Myth’”. *Science* Vol. 319, 414-415.

Zoller, U. (1993), “Are lecture and learning compatible: Maybe for LOCS, unlikely for HOCS.” *J. Chem. Educ.* Vol. 70, No.3, 195-197.

Dr. Bruce Dunham
Department of Statistics
University of British Columbia
2207 Main Mall
V6T 1Z4
Tel: 604 822 4997
b.dunham@stat.ubc.ca

Dr. Gaitri Yapa
Department of Statistics
University of British Columbia
2207 Main Mall
V6T 1Z4
gaitri.yapa@stat.ubc.ca

Ms Eugenia Yu
Department of Statistics
University of British Columbia
2207 Main Mall
V6T 1Z4
Tel: 604 827 5756
eugenia@stat.ubc.ca
