Implementing Specifications Grading in a Statistics Course

Eric M. Reyes

Department of Mathematics

USCOTS 2017

1 The Book

This workshop is based on Linda B. Nilson's book *Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time*. Nilson spends the first chapter exploring the motivations for a new grading system – namely, our frustrations with the status quo. Chapters 2 4 act as the foundational material for her proposed solution, specifications grading. These chapters discuss the use of course learning objectives, course design (linking objectives to assessments), and pass/fail grading. Chapters 5 and 6 introduce the essentials of specifications grading. Chapters 7 9 are peppered with course examples as well as testimony and research indicating both improved professor and student outcomes. Chapter 10 is a final justification for specifications grading. The book is littered with references to the education literature to substantiate her claims regarding problems with the current approach to grading and the components she stresses in specifications grading.

The book is a fairly easy read. I found that I very much resonated with the discussion in the first chapter. If you have experience in course design, Chapters 2 and 3 are primarily review. Chapters 4 6 really provided the meat of the text in my opinion. However, while they provide a nice framework, I was left with several questions about how this could be applied in a statistics course. Specifically, while the material throughout the text is presented in generalities, the majority of examples are from courses teaching humanities (literature, composition, etc.). I left feeling that if I were designing a statistical literacy course, I had a firm set of examples to follow. However, for a course emphasizing methodology, the examples presented may be disappointing. The last several chapters of the book were lacking as a result. Despite that, the text did motivate me to put this into practice, and I am slowly converting all my courses to this

grading style.

Our institution used the text as the topic of an interdisciplinary book club. The book club met regularly to discuss the text, and then each member spent a month putting the ideas into action either through redesigning a course or an assignment using specifications grading with the aid of course designers on campus. We met one final time to share ideas and discuss our progress. This was a great way to engage with the material presented in the book. That material serves as the foundation for this workshop. We will largely follow the outline presented in the book, discussing some key questions raised in reading the text. Specifically, how might we apply these concepts to a statistics course?

2 Motivation

Do you enjoy grading? If you are like me, the answer is a resounding no. But why? What makes it so loathsome? While I knew this was the least enjoyable part of my job, I was not capable of pinpointing a reason until recently. Here is an excerpt from Specifications Grading in which an instructor communicates her feelings on grading:

> *The passion I had for teaching was being killed by the dread I continuously faced with grading and worse, handing back the graded assignments. I found the after-grade encounters often antagonistic and more often related to winning back points than understanding the lessons to be learned from the experience. (page 6)*

What I have come to accept is that currently, grading is one of the most unattractive parts of being a faculty member. But, assessment is one of the most beneficial aspects to students. We have to ask is there a better way. Before we investigate a new system, it is worth critically considering the system we have currently adopted.

Worksheet

What does your current grading system do well?	What is broken about your current grading system?

Criteria for Evaluating a Grading System

Nilson identifies the following 15 criteria for evaluating a grading system. She argues that our status quo is not sufficient and argues we need something new. The last chapter of the book returns to these criteria to establish that specifications grading is a solution that meets these criteria.

1. Uphold high academic standards
2. Reflect student learning outcomes
3. Motivate students to learn
4. Motivate students to excel
5. Discourage cheating
6. Reduce student stress
7. Make students feel responsible for their grades
8. Minimize conflict between faculty and students
9. Save faculty time
10. Give students feedback they will use
11. Make expectations clear
12. Foster higher-order cognitive development and creativity
13. Assess authentically
14. Have high interrater agreement
15. Be simple

These are very big promises. It is natural to be skeptical that any system could deliver on these. However, after trying it, I am convinced that specifications grading, while maybe not perfect, does meet many of the objectives I wanted. Qualitatively, I have observed an improvement in student work, grading has been faster and less tedious, and my feedback has been taken seriously and created rich discussions between me and my students.

3 Course Design

There are many approaches and components to good course design. Essentially, good course design is about tying the objectives of the course to the assessments (exams, homework, etc.) given in the course to the grade students receive. Our goal is not to provide a comprehensive look at course design but provide a brief outline to sufficiently understand how course design will impact specifications grading.

Learning Objectives

In my opinion, the central component of all course design frameworks is well-written learning objectives (or outcomes).

> *A learning outcome is an observable ability or skill that a student is supposed to acquire by the end of a learning unit... (page 17)*

Emphasis is placed on a learning outcome being *observable*, or measurable. As an example, consider the following (poorly written) introductory statistics course objective (this actually appears as the official course objective at my institution for our introductory course):

Example 1. *(Poorly Written Objective)* Understand a sampling distribution and how it applies to making statistical inferences based on samples of data.

From an instructors point of view, how do we measure "understanding?" This leads to variability among instructors when doing team-based instruction. From a students point of view, how should they focus their studies for an exam? Does this mean they should identify the sampling distribution for a problem (multiple choice), compute probabilities for a sampling distribution (quantitative), or define a sampling distribution (short answer)? There are several things that might be meant by such an objective.

Example 2. *(Alternate Measurable Objectives)* Here are a few ways we might refine the above objective.

- **Define** what is meant by the *sampling distribution* of a statistic.
- **Identify** three models for the sampling distribution of the sample mean: bootstrapping, Central Limit Theorem, Student's Theorem.
- **State** the conditions required for applying the Central Limit Theorem: sample is drawn at random from the population; sample size is large; population has a finite mean.
- Given the mean and standard deviation for the population, **compute** the probability that the sample mean for a random sample of size n would exceed a specified threshold.
- **Discuss** the role of the sampling distribution when computing a p-value.

Certainly there are more objectives that could be listed. Notice that each of these is measurable. They define exactly how a student will be assessed on the knowledge and under what conditions. Well-written learning objectives can keep faculty on task throughout a term as well as clearly inform students of course expectations.

There are two levels of learning objectives: course-level objectives and unit-level objectives. Generally, a course has between five and ten learning objectives. Each unit in the course then provides several learning objectives that support the course objectives. Nilson relies on Fink's (*Creating significant learning experiences: an integrated approach to designing college courses, 2003*) six categories for categorizing learning objectives (as opposed to Bloom's taxonomy):

1. *Foundational knowledge* is mastering the basic information and fundamental ideas of the field or subfield around which the other types of learning revolve.
2. *Application* is practicing important skills and thinking about the knowledge in practical, critical, and creative ways, all of which make the other kinds of learning useful.
3. *Integration* is making connections among ideas, subfields, disciplines, aspects of one's life, and kinds of learning.
4. *The human dimension* is giving students new insights into themselves and others, and the human implications of other kinds of learning.

5. *Caring* is motivating students to adopt new interests, emotions, attitudes, and/or values related to the material and to want to learn more about it.

6. *Learning how to learn* is adopting self-conscious strategies to learn the material more efficiently and effectively and to learn more generally beyond the course.

As you progress down the list, Fink argues that you are achieving more advanced learning.

Once you have a list of course objectives, they should guide your assessments. You write homework and exam problems that reflect the objectives you have for the course. Students know exactly what to expect because you have provided them a list of objectives you expect them to meet by the end of the unit/course. The hard part is ensuring that you only write objectives you honestly intend to assess. For example, suppose your objective was

> **Comment** on the adequacy of a statistical method for addressing a given question of interest by **assessing** the conditions underlying the chosen method.

If we fall into a rut of only giving computational questions asking for a p-value, we will never really assess this objective. This objective requires, at a minimum, an essay question asking whether a p-value is valid to interpret. It might also be assessed by asking students to justify their chosen methodology in a course project. That is, it will require asking students to write, and it will require our being willing to grade a written response.

Worksheet 2

Type of Learning Objective	Learning Objective	Assessment(s)
Foundational Knowledge What key information is important for students to understand and remember in the future?		
Application What skills do students need to gain (analyze, evaluate, create, problem solve)?		
Integration What connections should students recognize and make among ideas within this course, with other courses, and in their professional life?		
Human Dimension What should students learn about themselves or interacting with others?		
Caring What changes/values do you hope students will adopt?		
Learning How to Learn What would you like students to learn about how to become a self-directed learner of this subject?		

Linking Grades and Learning Objectives

It may be somewhat clear how a learning objective can produce a meaningful assessment. But, linking course grades to the learning objectives is not necessarily clear. Nilson essentially presents three paths for linking these two components:

1. *More hurdles:* Higher grades reflect mastery of more material, even if that material is not necessarily more advanced. This generally links to accomplishing more learning objectives.
2. *Higher hurdles:* Higher grades reflect mastery of more advanced learning. Careful here that she does not mean more advanced topics (like a B for getting through 1-sample inference and an A for getting through 2-sample inference...that would be more hurdles). This reflects advanced learning (see Fink's six classifications above). This might mean, for example, that a B can be achieved by mastering computational problems but an A requires communicating the results through a written report (requiring completing a more advanced learning objective).
3. *Both:* Higher grades reflect mastery of both meeting more objectives and more advanced objectives.

As an example, consider a hypothetical 10-week statistical literacy course. The course, aimed at non-technical disciplines, provides an introduction to statistical thinking sufficient for engaging with data presented in the media and corporate reports. The course has the following learning objectives:

- Given a research question, **identify** the population of interest and the parameter of interest.
- Given a graphical summary, **critique** and **interpret** the graphic in the context of a research question.
- **Describe** the role of a sampling distribution in making inference on a population parameter.
- **Quantify** the level of risk they find acceptable when making a data-driven decision.
- **Appreciate** the role of statistics in making evidence-based decisions and **recognize** its limitations.
- Given a media article, **summarize** the strength of the evidence presented.

This hypothetical course is broken into four units:

i. Sampling Variability (2 weeks): building toward the notion of a sampling distribution, we discuss essential probability, numerical summaries, populations and samples, variability across repeated sampling.

ii. Data Visualization (2 weeks): elements of a good graphic and interpreting graphical summaries.

iii. Study Design (2 weeks): importance of student design on the interpretation of a study including observational studies, controlled experiments, and the notion of confounding.

iv. Confidence Intervals (2 weeks): incorporating variability into the estimation of a parameter with a focus on interpretation.

v. Hypothesis Testing (2 weeks): measuring the strength of the evidence against a claim with a focus on interpreting a p-value.

Consider the following grading scheme for the course:

Grade	Requirements
A	Requirements for D plus 3 additional Homeworks, 3 additional Reports, and Final Project.
B	Requirements for D plus 2 additional Homeworks and 2 additional Reports.
C	Requirements for D plus 1 additional Homework and 1 additional Report.
D	Pass Homeworks 1-2 and Report 1.
F	Failure to meet the requirements for a D.

This is an example of requiring more hurdles (more content/assignments) as well as higher hurdles (project) to get a higher grade. Notice that we do not have percentage grades associated with each assignment or the grade overall; this is a component of specifications grading which will be discussed as we move forward.

Worksheet 3 When mapping grades to collections of assignments, it is helpful to ask the following questions to get started.

What should every student who leaves my course be required to have *mastered*?

To what degree must a student master the material presented and still pass the course? Take the next course in the sequence?

What material is really optional but distinguishes awareness from expertise?

4 Essentials of Specifications Grading

Once you have a good course design in place, we can begin discussing how assessments are graded.

Pass/Fail Grading

While Nilson discusses some hybrid schemes which make use of components of specifications but still rely on a points-based system, her vision for a true specifications grading system requires that all assignments be graded pass/fail. To motivate this, consider the following excerpt from Nilson's book:

> *If you are like most faculty, you are sometimes unhappy with the quality of the work your students hand in, and you know they could do better if they just made the effort and took the time. Often enough to raise concern, they do not follow directions, answer the question, or complete the problems or writing task. Other times they do a shallow job, spending only minutes the previous night or just before class whipping out an assignment that should have taken them at least one or two hours. It seems they are not trying very hard. But why should they? We will accept the work to grade and will look for things to give them points for... (page 47)*

For as many fears of pass/fail as you may initially have, students will have more. There are lots of components to specifications grading, but grading things pass/fail is the hang-up for the majority of my students. The key to getting past this is to recognize that "pass" is not equivalent to "perfect." The criteria for successfully completing the assignment are defined by the instructor (discussed more in the upcoming sections). There are several benefits mentioned throughout the book regarding a pass/fail system.

1. Saves faculty time: we do not spend time deciding how many points each type of mistake is worth; we simply determine if it is a pass or not.
2. Increases student motivation: if students are to pass an assignment, they must work hard to master the material.
3. Improves quality of the work: students can no longer depend on producing enough disconnected thoughts to earn enough partial credit to pass. They must demonstrate mastery.
4. Mimics industry experience: for many of the professions our students are entering, either the work is right, or it is not. They will not earn partial credit from employers.

Specifications

For Nilson, grading pass/fail is not sufficient, it is only a component of the grading system. In her words:

> *What are we really assessing when we are grading pass/fail? We are assessing whether a student's work meets certain specifications – that is, one or more requirements that we set for a piece of student work. (page 56)*

She compares grading under this system to testing a piece of software. The software has certain specs that it is required to meet: it must run in a certain time frame, it must solve a particular problem, it must provide output in a specified manner. The software either meets the specs or it does not. If it fails to meet the specs, it will not be considered acceptable (requiring revision or abandonment). As faculty, we set specifications for an assignment. If the student's work meets the specifications, it earns a pass; otherwise, it merits a fail. There is no partial credit for meeting only some of the specifications.

For those that use rubrics to grade assignments, as Nilson explains, we can think of specifications grading as a rubric which only has a single level. A student receives a pass if and only if they have completed all specifications. However, we should demonstrate restraint as faculty; specs grading should not assess a work on every possible requirement we can conceive of. The specifications define what we consider to be minimally acceptable for the desired outcome.

Example 9. *(Article Reviews)* For my Biostatistics course (an intermediate course), students must review a series of scientific articles from the literature. For each article, I provide a score sheet which asks several general questions (What was the primary research objective? What was the parameter of interest? What statistical method was employed? etc). Here is the excerpt from my syllabus describing the criteria for a successful completion of the assignment:

> Successful completion of an Article Review entails a "good faith effort." Specifically, the following is required to satisfy a good faith effort:
>
> - Your score sheet, with comments, is uploaded by the due date.
> - At least one substantial comment is made regarding each category. A comment is substantial if, in the judgment of the instructor, an honest attempt was made to critique the paper (positively or negatively) **and** evidence from the paper is cited.

> Note: your opinion of th earticle does nto factor into your grade (you can love it or hate it and still receive credit for the assignment).

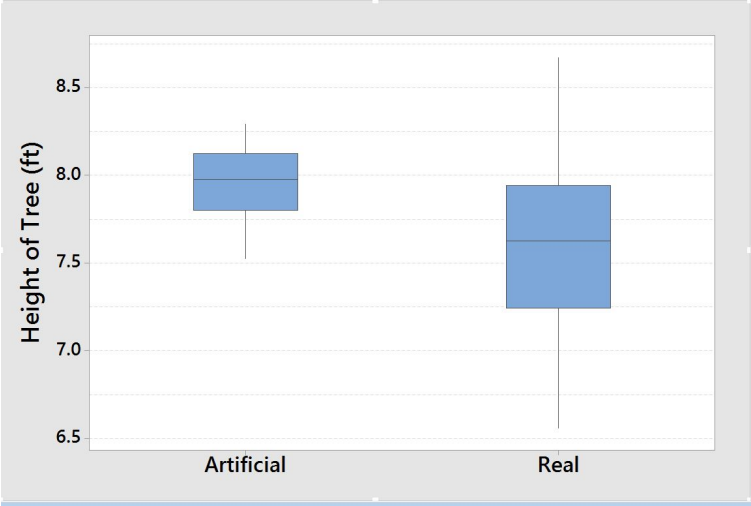
Note that the criteria are not all-encompassing. I do not look at grammar for this assignment. I do not even evaluate whether they correctly address each question or category. I am solely looking for evidence that they actually read the paper. This coincides with the primary learning objective I am aiming for with this assignment:

- **Assess** the strength of evidence presented by a scientific publication in addressing a research question and **provide** constructive feedback for improving a study.

Most of the "learning" I expect will take place during the in-class discussion of the article. This critique is mostly to ensure that students are prepared to engage in a meaningful in-class discussion. This has boosted participation in the article discussions (I did not require written review in the past), does not really add to my grading load (it takes only a few moments to look through all assignments), and encourages students to take notes prior to our in-class discussions. Clearly, the specs for a course project would be more encompassing; specs are tied to the instructor's goal for an assignment.

I will note that exams are fundamentally different than a writing assignment (be it a project or otherwise). Embedding a classical exam in this framework might require us to think carefully about how we are assessing problems and the role an exam plays in our final grading scheme (remember, we are not weighting exams and homework assignments anymore).

Worksheet 4 *The Grinch plans to take the Christmas trees from each home in Whoville in his attempts to spoil Christmas. In order to ensure he has room on his sleigh, he is interested in determining if there is a difference in the height, on average, of Artificial Christmas trees and Real Christmas trees. He randomly selects 20 homes which had an Artificial tree and 20 homes which had a Real tree (for a total of 40 homes). For each of the 40 homes, he measured the height (in feet) of the tree. A graphical summary of his data is available below:*



Suppose we were to conduct an ANOVA to address the Grinch's question of interest. Would we expect the corresponding p-value to be large or small?

Develop a set of specifications for grading this question. How can you clearly express to students what is required to successfully address such questions?

Example Work for Students

Specifications grading requires faculty to be very clear when communicating our standards.

> *We [the faculty] take on the responsibility for providing students with all the information, instructions, and models that they need to succeed. Our specs have to be crystal clear, more so than the brief descriptions typically found in multilevel rubrics. This means taking the guesswork and mind-reading out of the student experience... (page 58)*
>
> *In the current grading system, we are not expected to lay out the template in such detail, so we don't. And our students do not always pay careful attention to the sketchy (to them) instructions they get. Besides, they can bank on partial credit for just making a weak attempt. No wonder the quality of work they submit varies radically. We are willing to accept, "settle for," and give partial credit to work that falls short of what we define as good. (page 61)*

Our specifications can be very detailed (focusing on the mechanics of an assignment) or very broad (focusing more in the instructions and less on the response). Either way, they need to be clearly written so that there is no doubt in a student's mind what is expected.

The fear of course is that this produces a culture of students asking "will you check this for me and tell me if it is right?" The clearer the specs, the less likely this is to occur. Students should have a checklist that they can go down and determine if the assignment is adequate. To alleviate this, strongly consider giving examples of both acceptable and unacceptable work. Further, I recommend that you annotate these examples to specifically describe in what ways the work met or failed to meet the specifications.

Substantially more work is required on the part of the faculty member prior to the course running to put examples in place. This is especially true when constructing examples of projects. For example, I provide 3 examples: one illustrating a project which meets all specifications, a project which nearly meets all specifications and areas where it could be improved upon, and a project which fails to meet specifications. Each is annotated to provide students with information on how the specifications were met/violated.

Flexibility/Tokens

Remember, specifications do not need to reflect perfection. Indeed, Nilson states than in general, an appropriate level for a "pass" should reflect B-level work (or 80%). Again, this threshold is up to each faculty member. Your goal is to set a specification to achieve the minimally accepted level of learning by the student. Even for students who are familiar with this type of grading system (and most will not be), grading pass/fail can be intimidating. Nilson suggests we can add flexibility through the inclusion of tokens.

Essentially, Nilson describes setting up an economy in the classroom. Students are given, or have the opportunity to earn, tokens throughout the term. Each token has a value for "purchasing" a commodity. For example, a token might entitle a student to a re-submission of a homework assignment. Or, you might allow two submissions of the project. The use of tokens can reduce stress on students regarding their grades.

Bundles and Modules

Remember that in Nilson's vision for specifications grading, we no longer weight sections of the course to develop a course grade (however, she does discuss such "synthetic" approaches briefly in the book). Instead, grades are determined by completing a specified collection of assessments. To aid in this process, Nilson introduces modules and bundles – collections of assessments which are graded pass/fail based on specifications.

> ***Modules*** *are units that are sequenced during the term, and* ***bundles*** *are units that students can complete in any order...Which student products to include in these units of work are among the basic course design decisions an instructor makes when implementing specs grading, and they can neatly tie into learning outcomes. (page 69)*

Below is an example of how we might structure an introductory course. Note: I am not necessarily recommending this; it is simply provides a short example.

Example 13. *(Introductory Course)* We might construct an introductory course with five bundles:

- Bundle 1: One-Sample Inference about the Mean
- Bundle 2: Inference about the Difference in Means from Two Independent Samples

- Bundle 3: Inference about the Difference in Means from Paired Samples
- Bundle 4: Analysis of Variance
- Bundle 5: Simple Linear Regression

Each bundle consists of three assessment items: a homework assignment, an exam, and a case study. All three items must be successfully completed in order to complete a bundle. Successful completion of each assessment is described further in the syllabus. Grades are earned according to the following:

Grade	Requirements
A	Complete all 5 bundles
B	Complete Bundle 1 and 2 of the remaining bundles
C	Complete Bundle 1 and 1 of the remaining bundles
D	Complete Bundle 1
F	Fail to meet the requirements for a D

Care would need to be taken regarding when bundles were due so that all work is not submitted at the end of the term. One of the most difficult things to do when determining the grading scale is to think through "what if" scenarios regarding student's submitting substandard work. While this should not occur too often, it will happen. If a student aiming for an A struggles with completing Bundle 1 above for example, but successfully completes Bundles 2-5, does he or she deserve to fail the course? These types of questions really get back to good course design and asking ourselves what is required to successfully pass our course. I have found that the use of modules and bundles is easier to conceptualize for an intermediate or advanced course, especially those taken online.

Worksheet 5

Bundle/Module	Content & Assessments	Supported Objectives

5 Examples

The examples presented in Nilson's book tend to come from non-STEM courses. I found it difficult to apply these to the courses that I teach. I have made available relevant portions of the course syllabus for three courses I currently teach under this specifications grading. They each have pros and cons. But, hopefully they provide some idea on how a course could be structured using these ideas. The Biostatistics course and the Statistical Computing course are truly specifications grading. The Introductory course is a synthetic course (assignments are graded pass/fail, but each is linked to a number of points that are earned).

If you are just beginning to explore specifications grading, it is natural to have more questions than answers at first. Remember, the key is that you, as the faculty, will define what expectations must be met by students to merit a particular grade in the course. You define what it means to successfully complete an assignment. This is all done before-hand. Once class begins, students choose which grade they want to earn. You are simply playing the role of helping students get to their desired grade.