AC 2012-3873: TEST PREPARATION AND TEST QUALITY ASSESSMENT:
WHAT I WISH SOMEONE HAD TOLD ME IN THE BEGINNING

Prof. David B. Meredith, Pennsylvania State University, Fayette

David Meredith is an Associate Professor of general engineering with more than 30 years of teaching experience at Penn State, Fayette, the Eberly campus. He teaches both engineering and engineering technology classes. He is a registered Professional Engineer and active in ASHRAE, ABET, and NCEES. He has received numerous awards from the campus, college, university and other organizations for excellence in teaching, scholarship, community service, and advising.
Test Preparation and Test Quality Assessment –
What I Wish Someone Had Told Me in the Beginning

Abstract
How does an instructor prepare a test and feel confident that it is fair, balanced and the correct length? That issue worried me as I entered the teaching profession three decades ago. After a decade of writing multiple choice questions for the Professional Engineering exam and two decades of national involvement with the ABET accreditation process, it is time to pass some of the “tricks of the trade” along to the next generation of new faculty members.

As a starting point for this discussion, the assumption that course outcomes have been linked to not only Program outcomes but also ABET outcomes will be reviewed. While discussing the selection of an appropriate evaluation method, I will also discuss some of the topics that do NOT lend themselves to timed tests. Suggestions on alternative methods of evaluation for those topics will be offered.

This paper will focus on how to prepare good True/False and Multiple Choice questions to test knowledge and skill. It will also present a simplified version of the method psychometricians use to evaluate the quality of these types of exam questions for both difficulty and discrimination. Since many engineering and technology questions involve calculations, a test evaluation method that works for me will be discussed. This method is a good first step to demonstrate course improvement to ABET. Since much of how we teach is driven by ABET criteria and the need for continuous quality improvement, this experience should prove valuable to new faculty who may never have prepared for an ABET visit.

Introduction
When I learned that I would be starting a teaching job at the end of my last semester of graduate school, I panicked. I had done some tutoring, so I was comfortable with the basic learning process. But I had no clue on how to prepare exams or evaluate final grades. Over the years, I have developed some techniques that work for me. Hopefully they will work for you.

Developing Course Outcomes
The first step is to identify the specific concepts and skills that represent successful completion of the course and express these as course learning outcomes in proper format. ¹ This step is also required as part of the ABET CQI process. If the course has already been taught, the course outcomes should already be included in the existing course syllabus and should be reviewed as part of your course preparation. I have always viewed this process as deciding what will be covered on the final exam before I teach the first lesson. Keep in mind that your course outcomes are always a work in progress. In fact, it is a good idea to review and update them at the end of each semester, while everything is still fresh in your mind.

Philosophy of Testing
From my perspective, there are three purposes of a test: (1) the students learn what they don’t know, (2) the instructor determines where the course can be improved and, oh yes, (3) we have to assign a grade at the end of the semester. In my opinion, that is the order of importance. Too often our students (and sometimes the instructors) view the exam as simply data to determine the final grade. In fact the exam provides key information to both parties on the learning (and teaching) progress that has been made. The exam also helps identify holes in the students’ understand or skill development while there is still time to improve.

Perhaps the wisest advice that I ever heard a practicing engineer relate to students was “Make as many mistakes as you can while still in college and learn from them. It will cost you some points, but you will still graduate.
Make that same mistake in my office, and it could cost a million dollars…or possibly somebody’s life.” I share that advice with every class I teach.

One issue that Engineering instructors seldom deals with is Test Anxiety^2 in students. Our assumption is that students should be “over it” by the time they get to our courses. And if the problem still exists, other faculty members should take time out of their class to “fix” it. But there are some simple changes that we can make to help reduce student angst toward tests in our classes.

The traditional European learning process focused grading on a single exam at the end of the learning experience. Everything rode on that one performance. The learning style of today’s students (cram the night before the test^3) is more conducive to more frequent exams covering less content. Today I usually give three midterms plus a comprehensive final in each of my technical courses. In one course, the final grade is based on the final plus eight quizzes with no midterms. While this may sound excessive, remember that each quiz only covers a couple of topics. So it is only a couple of questions on a few concepts and usually only takes about thirty minutes of a class period.

One advantage of this method is that students get an early opportunity to learn my style of questioning. We all have our own way of phrasing things and asking questions. There is quite a bit of student angst (especially for students in their first few years of college) to try to “out-psych” the professor’s method. Giving the students practice in how to extract the important details from your problem statements helps alleviate this extra pressure. I traditionally distribute last year’s comparable exam to help students prepare for their exam. Their exam will cover the same material, but the order of the questions will change as well as the unknown they are looking for. For example, to test their ability to apply the equation “F=ma”, the practice test might ask for “m” given “F” and “a” while their test might ask to solve for “a” given “F” and “m”. The first test might involve two cars and the second test will look at two billiard balls.

Since I only see recent high school graduates, I consider it part of my responsibility to improve the problem solving skills of the students and (in many cases) to break some of the bad habits that they may have learned in high school. It amazes me every year how many engineering students have forgotten the formulae for volumes and surface areas. So they get some practice early in the semester on homework and entrance quizzes to ensure these basic equations are fresh in their memory. The students that I see are usually weak in converting units. So they quickly learn (because I tell them repeatedly) that if they need the velocity in m/s for the equation they are using, they can expect the velocity to be given in km/h. They need the practice converting units to those needed in the equation before they start solving the problem. The key to solving both of these issues is for each student to develop a “crib sheet” with key equations, conversions, concepts and terms. I often ask alumni who are 20 years out of school if they still have their crib sheet (“yes”) and when they last used it (“last week”).

I have also found them to be weak on thinking in multiple process steps. So through homework problems and practice tests, they learn to calculate the volume of a container given the dimensions, convert to mass using the density, and use that mass value in the “F=ma” equation. The final skill that I try to develop in these budding designers is the ability to filter extraneous information out of the problem statement. There seems to be an unwritten rule that the student’s solution must use every value given in the problem statement. By including the viscosity as well as the density in the problem statement, you can learn some creative student physics on how the values mesh together into an answer that you have never considered before.

It would be cruel to include all of these complications on every test question. So I generally limit their use to one or at most two questions on each exam. And I make sure I tell the students to expect some twists before the exam. Then I remind them that they are allowed to think during the exam. Too many times it seems they are so anxious to start punching numbers into their calculator that they forget to consider why they are doing the steps.
Some topics are just not conducive to a timed test. When teaching Thermodynamics, students demonstrate their ability to accurately interpolate using the steam tables on the first test. After that, the hardest value to extract from the tables during a test might be to average between two tabled values. I see that they get plenty of interpolation practice in their homework. But for their later tests, I want to see if they know what to do with the values, more than if they can just determine them. Soft skills are better demonstrated through team projects and written lab reports. To test the students’ power of observation, I will devote one lab to tracing water lines in a laboratory and discussing related issues (e.g., Was there a floor drain?). To get them thinking “out of the box”, we will discuss alternative methods to verify an instrument reading. I often tell the “Angels on a Pin” story.

To strengthen their skill at estimating values, we will count ceiling tile or concrete blocks to determine the dimensions of a classroom. None of these traits could be demonstrated on a timed test, so get creative on how students can demonstrate those estimating skills other than through traditional testing.

I try to build the exam well ahead of the exam day. I let it rest overnight so I will read it with fresh eyes when I time the exam length. This ensures that I am reading the exam as it is written – not on what I planned to say. Quite often I find glitches which can range from conflicting values between the text and the graphic to missing parameters needed to solve the problem. The length of the exam is critical. The instructor should be able to complete the exam in 25% to 30% of the time allotted for students. So a 50-minute exam should take the instructor 12 to 16 minutes to complete. If it takes over 18 minutes for you to complete, you need to eliminate some calculation steps in a long problem, omit one or more questions from the exam, or allow students to choose one or more problems to skip. (If they all skip the same problem, there is a message for you about their confidence in that topic!)

Open-Ended Questions

Most engineering and technology test questions involve the classic “story problem” that almost all students loathe. Since a picture is worth a thousand words, I generally try to include a graphic with the problem statement. It is often easier for students to get a sense of the problem visually than it is for the instructor to write a complete verbal picture that all students will interpret correctly. If you are asking for the leaving velocity, indicate “V = ?” under the vector arrow so they know what they are looking for. The main caveat is to be sure that the data in the graphic matches that in the text – especially if you adjust the values as part of the test development process.

One test condition that I try to avoid is what I call “double jeopardy.” That is when the student must successfully solve for one answer using a recently-learned equation and then use that answer in a second equation. I have never figured out a clean method to fairly grade the second question if they miscalculate the first answer. My remedy is to provide the middle answer, then ask one question going forward and a second going backward. For example, in a flow situation, I might give the velocity, then ask for the specific kinetic energy (going forward) and (given the diameter) back out the volumetric flow rate. Note that this issue usually involves the application of newly-acquired concepts. Any core concepts (e.g., ideal gas laws) or material tested in a previous exam should not be a problem for students on the current exam. In other words, multi-step problems should involve both old and new material.

I assign the point values for each process step before looking at the first student’s exam. Generally I start with a 30/40/30 balance. The student gets 30% of the question value if they state the correct governing equation and correctly solve it algebraically for the missing value. They get about 40% of the points if they substitute the correct values for each variable – with the correct units. The final 30% just proves they know how to use their calculator. But if they forget the units or include inconsistent units, they lose some of these points. The final step to successfully completing an engineering calculation is to “reflect” on the answer to be sure it makes sense. (“The cup of coffee has a mass of 247.38 kg, because that is what my calculator shows.”) I have tried unsuccessfully to incorporate that step as part of this final 30% by offering 5% if they write “ok” beside their
answer to demonstrate they have completed this final reflection step. Some students thought they deserved the points because they had written “ok” without showing any work. I think they missed the purpose of the points.

I do try to provide feedback directly on the test to show the student where they went astray. Sometimes a student is so lost that it is easier to just write “We need to discuss – see me after class.” It is hard to correct their thought pattern if you cannot even figure out what they were thinking. Occasionally the whole class messes up a question. When that happens, I print out a correct solution for that problem that we discuss in class. At the bottom of the page is a new (similar) problem for them to solve and submit to recover some of the points lost on that question.

Occasionally a student will question the number of points they received on a question – they obviously deserved more (in their opinion). For those students, I always start the conversation with “Well, you have the wrong answer, so your starting point is zero points. Let’s talk about whether you deserve any points at all on this question.” Once they look at it that way, the students have always seemed to be grateful that they got some credit for a wrong answer. But I also show them that on my answer key the point distribution is shown and their points match up with my rubric.

Evaluating the Performance of Open-Ended Questions
Once you have scored the test, how do you evaluate what the students learned and where they still need to improve? The following technique was suggested to me by a former TAC of ABET chairman. I have been using it for the past several years, and find it to be a useful tool. It can also become an example of your CQI process for your next ABET visit.

As soon as the tests have been scored, select a representative sample (e.g. 3 high scores, 4 middle scores and 3 low scores). Set up a table with a column for each student and a row for each scored question. For each question, record the evaluation points (or deduction points, depending on which metric you use) for each question. Repeat the rest of the columns for each student’s test. The sum of each column represents the score for each student. The average for each row in the table represents the class performance for each question. I normalize the scores to make it easier to understand. For example, if the ten students in the sample totaled 40 points on a question worth 5 points each, they have earned 40 of 50 possible points collectively – or an acceptable score of 80% on that question. But on the next question, the average might only be 20%. That means there is a problem: either I didn’t teach the topic as well as I thought I did, they didn’t understand the topic as well as I thought they should have, or I didn’t ask a very clear question. So I make notes beside the lowest scores to record the type of student errors made – math error, wrong data from a table, wrong equation, forgot a step, etc. To complete the continuous quality improvement loop, I also record a comment in the relevant section of my lecture notes on how I can improve the students’ learning related to that skill or concept for next year. That might involve more examples, more practice problems, or just more class time on that topic. Always remember that the first letter of CQI stands for continuous. You cannot fix everything simultaneously. So use this “triage” method to determine where the greatest gains in student performance can be found for the least investment in your time.

While open-ended exams generally take less time to assemble, scoring them is a slow process and arguably subjective. Conversely, Multiple Choice (MC) Tests take longer to assemble, but only a few minutes each to score. The larger the class, the more sense it makes to consider alternative testing strategies. For knowledge-based information, MC is better than fill-in-the-blank. (Although as a rookie instructor, I kept track of students’ common wrong responses – they make great distracter answers on a MC test.) So a typical test might be a blend, mixing MC questions to test knowledge and Open-Ended questions to test computational skill. But is it possible to test students’ computational skills with MC questions?
Multiple Choice Questions

The three parts of a MC question are the root statement (called the stem), the correct answer and three (sometimes more, seldom less) alternative incorrect answers (called distracters). Obviously the stem must include all information needed to clearly discriminate the correct answer. Arrange word answers alphabetically (or numerically) to randomize them. Keep a consistent length for all responses – a response with extra details is a dead give-away for students who know engineers love details. Try to avoid “none of the above” or “all of the above” – psychometricians claim they are also give-aways.

Initially, I did not think MC questions could test a student’s ability to solve mathematical problems. But in 2000, I got involved with writing questions for the Professional Engineer exam for NCEES, and they only use Multiple Choice questions to test competency. Today, I am a believer that it can be done. But it does take more time to develop good questions.

When using MC questions to test skill, it is even more important to include all of the information needed to solve the problem in the problem statement. Include a graphic if possible identifying known conditions, and the location and type of value being determined. Minimize the number of calculation steps – it is difficult to offer partial credit using MC questions. So the test time students invest for each MC question should only require a few minutes. On the PE exam, we assume an average of 6 minutes per test question. However it can make sense to include two or three MC questions related to the same stem. Students have already processed the basic information, filtered out unneeded data, and possibly completed some unit conversions. Use that situational familiarity to your advantage and ask a second question using the same stem.

The final line in the stem should include the expected units on the numerical answer and be phrased similar to: “The mass flow rate (kg/s) is most nearly:” This provides for rounding error and (heaven forbid!) the possibility that the answer is above or below the range of answers provided.

Arrange numerical answers in either ascending or descending order. All answers must be consistent in the number of significant digits. (Avoid 0.1, 1.0 6.75, 10.) To provide for rounding error, answers should be no closer than 10-20% of each other. (Avoid 98.2, 98.8, 99.1, 99.5 – unless the calculation method can discriminate among these values.) And most importantly, use distracters that weak students are likely to generate. Tracking mistakes on previous similar questions is a great way of determining these errors.

My experience is that students commonly forget key scale conversions (e.g., not converting to absolute temperature or pressure), forgetting key unit conversions (e.g., km/h to m/s or kilograms for grams), make silly math errors (e.g., \(10 - [-2] = 8\) instead of 12), forget to square or square root terms (or go the wrong way) and randomly interchange diameter with radius.

True / False Questions

Students have often told me that true/false questions are the hardest type to answer and I agree with them. As hard as we try, T/F questions with “always” or “never” will include an unexpected exception. If we use the terms “usually” or “seldom” students can get argumentative. It is easy for students to focus on the wrong issue or to get sidetracked based on a creative interpretation of what the question is testing. To help guide students to respond to what I am looking for, for each question I select a “trigger” word and highlight it by making it ALL CAPS or bold. Before the test cycle begins, we practice replacing the trigger word with the opposite to see which makes better sense. Consider the example: “When laying sod, put the GREEN side up.” Would it make better sense to replace “green” with “brown”?
I also tell students that I reserve the right to modify a question that they may have seen before. So the above example might show up as “When laying sod, put the green side DOWN.” That tests the same knowledge, but students still have to think about the situation before responding.

**Matching**

Matching questions are basically Multiple Choice questions with more than four possible answers. I find them most useful when testing knowledge of multiple characteristics for several alternative concepts. Put the list in alphabetical or numerical order to randomize. To ensure students don’t assume they can guess the last answer by default from what is unused on the list, my instructions state: “You do NOT have to use all answers, but you CAN use an answer more than once.” Then I make sure that they never come out even.

**Evaluating the Quality of MC, T/F and Matching questions**

There are two metrics that psychometricians use to measure the quality of a True/False, Multiple Choice or Matching question: difficulty and discrimination. Difficulty ensures that the question was not too easy or too hard. Discrimination indicates that the question correctly separates stronger students from weaker students.

After you have scored an exam, select representative samples from the top and bottom of the scoring range. Ideally, these two cohorts should represent the top and bottom quartile of the data. That is easily done if you use computerized grading. But when I hand-score tests, I usually select up to ten tests from each extreme. To determine the Difficulty, add the total number of correct answers and divide by the total number in the combined samples to get the average score. The Difficulty value should be between 50% and 90%. Obviously you don’t want it to be too easy or too difficult. If it drops much below 50%, guessing becomes as efficient as knowing the facts.

To determine the Discrimination, subtract the number of correct answers from the Lower cohort from the number of correct answers in the Upper cohort. Then divide by the half the total number in the sample. For example, if the Upper and Lower cohorts scored 8 and 4 respectively (out of ten each), the Discrimination value is \((8 - 4) / 10 = 40\%\). The Discrimination value should be between 30% and 70%.

A “perfect” question (at NCEES, they are called “equators”) will meet both criteria. If just 25% of your questions rise to that level, you have written a great test. I have been told by psychometricians that grades can be assigned with as few as 4-6 “perfect” questions. The engineer in me wants to err on the side of safety, so I would recommend at least a safety margin of 2:1 (ten questions minimum) for a knowledge-based quiz. At NCEES, the exam specification requires each exam to include 25% equators, 25% new questions and the rest can be recycled questions. But then, they are professionals with both budget and staff. For the rest of us in the “real” world, keep a file of your equators and try to add some new ones each cycle.

I modify the above data collection process into two steps. Immediately after scoring a test, I select my sample cohorts and record the number of correct answers for both cohorts for each question directly in the margin of my answer key. That allows me to return the tests to the students by the following class so they get more timely feedback. To process the data later, I first do the math to determine the range of totals (difficulty) and differences (discrimination) needed for each question. If the difficulty level for each question is in the 50-90% range, I use a highlighter on the Lower cohort number for that question. Occasionally a question will have a negative correlation (the lower cohort scores better than the upper cohort). These questions also get an extra dot of the second color in the margin.
The final step is to summarize the data by counting the number of questions meeting the criteria in each category. I draw a two-circle Venn diagram (since the data is for my personal use primarily). The number in the left circle represents the number of questions meeting the difficulty criteria. The number in the right circle represents the number of questions meeting the discrimination criteria. The number in the intersection of the circles represents the number of questions meeting both criteria. To the left I record the number of questions that were too hard – those represent knowledge that must be better explained next year. To the right, I record the number of questions that were too easy – that might be knowledge that you may not have to test next year. Or perhaps you can reduce the time used to present the “easy” topic to allow more time for topics that did not test well. Just be sure to capture this data where it will help you improve the course next cycle. Below the Venn diagram I record the number of questions that the better students got wrong more often than the weaker students. I try to interpret what might have caused that reversal – did the better students overthink the problem? Did they make other assumptions?

**Returning Tests**

My goal is to return scored tests to students within one week following the exam. Students who miss the exam have that interval week to make it up or come up with an extremely creative and entertaining excuse. Once the exams have been returned, these students are more likely to see a classic exam from a previous year, a situation that is never to their advantage.

When returning the first exam of the semester, I take class time to explain my test grading philosophy. I list the test scores on the board from highest to lowest and indicate the median score. In a perfect world, the high score will be in the high 90s and the average score will be in the mid 70s. That might have happened a couple of times for me in the past 30 years. More commonly the high is like 82 and the average is 63. Unless you know that the smartest student in the class has just over a 3.0 GPA, the test was either too long or too hard. Since students like to know how well they are doing, I try to show them where the letter grade cut offs should be. The easiest method is to add a fixed number of points to their raw score (in this example, adding 10 points might be appropriate) and using a 90% / 80% / 70% / 60% for the A/B/C/D ranges. I clearly explain that I do NOT assign letter grades to tests and only record raw point scores. I do point out to students with the lowest scores that they will need to make some extra effort to avoid failing or having to repeat the course. On subsequent tests, I only list the highest score, the average and the lowest score. I repeat the verbal warning for the weaker students each test if the lowest score is well below the “C” level.

So what does it mean if the test scores are significantly below what you expected? Does that mean this year’s class is exceptionally weak? Or did they study the wrong material? Perhaps they never tried the homework problems you assigned – the test was modeled after those examples (or at least it was in the instructor’s opinion – student views may be different). Maybe they just didn’t follow my perfect lectures because I am sure we talked about all of these concepts in class…or was that last year? Or could it be that my perfect test left something to be desired?
Seeking Student Feedback

As part of my preparation for the final exam, I distribute a survey listing the course objectives as shown on the original syllabus. I ask the students to rate their confidence in each item using a “+” (good), “0” (acceptable) or “−” (weak). They can also use a “++” or “−−” for the extremes. Obviously, this is just a Likert scale using symbols, but my students seem more comfortable with it. I also ask them to select one outcome that they would have liked to spend more time discussing (marked with an “M”) and one topic that they felt could be effectively covered in less time (noted with an “L”). There is also a “free response” section on this survey sheet. I do make a verbal (and honest) point while handing out the survey forms that this survey will in no way affect their grade, and that I will not even look at the data until final grades are submitted. If students are agreeable, we spend about 15-20 minutes discussing ideas that could improve the course. I find this feedback invaluable. I document the changes made (based on student comments) for my next ABET visit. After grades are submitted, I do a quick spreadsheet on their survey responses and calculate the average for each outcome. The highest few averages I highlight in green and the lowest in red or yellow (depending on how far below normal they are). Continuous Quality Improvement does not mean you have to fix everything at once. I try to think of ways to improve the lowest rated outcome each semester – or at least develop an explanation on why students rated it poorly. That feedback is also transferred into my notes for the next course cycle.

When all else fails, if you cannot identify why students are not performing up to your expectations, ask them. You can do this anonymously using “minute papers” or by asking your TAs for what they have heard. You can also ask students who ask for extra help during office hours. You might also be able get a peer review by having a colleague in the department take a look at your exam.

Conclusion

“You cannot improve what you cannot measure” is a mantra that is often heard in engineering discussions. So it is with the quality of testing. While it is possible to use the techniques discussed in this paper to demonstrate that a test covers the correct material, is reasonable challenging and about the correct length, constructing that “perfect” test still requires some artistic and creative abilities. By incorporating a couple of nuggets from this paper into your testing method, your students will improve their testing-taking abilities, and you will be more confident that your tests are more valid.

References

(4) “The Barometer Fable”, http://www.lhup.edu/~dsimanek/angelpin.htm