WHAT’S GOOD ENOUGH?

Setting Standards

Monica Stitt-Bergh
Assessment Office
University of Hawai‘i at Mānoa

Workshop at the Association for the Assessment of Learning in Higher Education
2014 Conference

Abstract: 100 is a good score. Or is it? A score of 100 doesn’t mean anything on its own. Standards provide the context or comparison that gives a score meaning. They help us interpret assessment results and figure out how the results can be used to improve teaching and learning. In this workshop, participants will learn standard-setting methods and practice them. In addition, participants will learn ways to set targets and tips for facilitating standard-setting sessions on their campus.

Handouts & Files
(1) Outline
(2) Performance Descriptions
(3) Student paper
(4) Written Communication VALUE rubric
(5) Recording form
(6) Multiple-choice exam standard-setting activity (Political Science)
(7) Facilitation Tips
(8) Excel file: calculating the “cut score”
Agenda

- Welcome
- Terminology, Purpose, Key Concepts
- Standard-setting Methods
  - Activity: Rubric-method
  - Paper Selection Method
  - Activity: Multiple-choice exam
- Set Targets (Benchmarks)
- Q&A
Today’s Outcomes

- Understand the purpose of standard setting
- Become familiar with standard-setting methods
I am one of two faculty members who run the Assessment Office at the University of Hawai‘i at Mānoa. Our office’s mission is to improve student learning through program and institutional learning outcomes assessment.

University of Hawai‘i at Mānoa is a research university: a land-, sea-, and space-grant university with 2,000 faculty members and 20,000 students. We are an indigenous serving institution, built on native lands and concerned with preserving and promoting the host culture.
I grew up near Lake Michigan and I am a graduate of the University of Michigan with a degree in English. I’ve lived in Hawai‘i for 25 years and have worked at the University of Hawai‘i since 1993.

I have a master’s degree in Rhetoric and Composition from the University of Hawai‘i Mānoa English Department and my doctorate is in educational psychology with a specialization in college-level writing assessment. Within ed psych, I’m a fan of Vygotsky and a historical-socio constructivist view of human learning and development.
I enter this workshop assuming that you’re pro-assessment. This is a major reason I’m happy to be here at AALHE. I’ll be preaching to the choir if I talk about the usefulness of assessment. I don’t have to deal with resistance to assessment. I don’t have to let you vent or talk about your fear of assessment. So thank you very much!
Standard setting answers the question, “what's good enough?” because by themselves, scores have very little meaning.

For example, do these scores and results demonstrate student proficiency?
On my campus, these scores are “good enough.” The CLEP and AP exam scores get students college credit and the student has met expectations for the Speech department.

It’s important to know that the standard setting process to determine what scores mean, requires human judgment. It cannot be done by a computer applying a mathematical algorithm. The need for human judgment is one reason why I believe standard setting processes are valuable to assessment leaders on campus: it values faculty members’ judgment and expertise.
I have four terms to define to help make sure we're on the same page.
First, OUTCOMES specify what students are expected to know, do, and value. In K-12, they use the phrase “content standards” and some people prefer the word “objective.”
Standards define levels of performance on the outcome so decisions or classifications can be made.

Because “standard” has many different layman’s definitions, when I work with faculty on my campus, I often say “setting performance expectations” and “establishing a cutscore” on the rubric or exam. “Cutscore” seems to be a word that faculty understand.

My tip is that when you’re searching the literature for more information, use “standard setting” in google scholar and library databases and when you’re working with faculty, use “expectations” and “cutscore.”
I’m defining benchmark as a reference point by which a program or campus can measure themselves for the purpose of comparison.

This example sentence should help clarify.
To set standards, you need to determine how many performance categories you’ll place students into. For example, two categories are the simplest: meets or does not meet.

HANDOUT: Performance descriptions based on the Degree Qualifications Profile, 2 categories
2. PURPOSE

Why set standards?
Purpose: Why set standards?

Scores by themselves have no meaning.
Once scores have meaning, this is what you can do with the standards.

1. Clarify performance expectations: we will always have students whose knowledge and skills vary. It’s helpful to students, faculty, support staff to share a common understanding of the performance expectations in concrete examples or terms.

2. Motivate greater level of achievement: though the clarification of expectations, students can be motivated to greater levels of achievement. They will have a clear understanding of what it means to meet the minimum level and to meet a higher level.

3. Allocate resources: once standards are set and the program determines the number of students who meet and who do not meet the expected performance level, resources can be allocated differently as a means to increase the number of students meeting the expectations. E.g., resources for tutoring or lowering class size.

4. Decide whether to award a certificate, license, college credit: high stakes testing organizations use standards to determine whether a person receives a license.

QUESTION for YOU: In regards to program or institutional learning outcomes assessment on your campus, which of these make sense? What might the purpose of standard setting be on your campus?
On our campus, program assessment is used for program improvement. So our primary purposes are to clarify performance expectations and to motivate greater levels of achievement. Over time, we advocate for change in resource allocations. We do NOT use program-level assessment activities for high stakes decision making.

IMPORTANT: Setting standards is political and emotional. I encourage you to carry out standard setting as part of program assessment, and when you do, be sure to explain and reassure people how the expectations/cutscores, will be used. Please address their fears first: e.g., standards will not be for individual student graduation decisions, not to evaluate individual professors. Used only for clarification of performance expectations and to motivate greater levels of achievement among students.
Finally, the main reason I recommend standard setting to you is because it can increase faculty engagement in assessment activities.

My philosophy rests on the belief that faculty are in the best position to help assessment reach its goal of program improvement. I believe that the closer the assessment activities are to faculty and students, the greater likelihood that assessment will happen and results will be used to guide program decision making.

Because of my philosophy, I am always interested in strategies to get faculty doing assessment and that lead to an observable product. Standard setting is perfect because it has a set of established steps, it takes a close look at student work and accomplishment, and leads to product: a performance expectation or “cutscore.” It is an excellent means to move faculty from abstract conversations to concrete knowledge and decisions.
3. KEY CONCEPTS

- Informed judgment
- Borderline student
The first key concept is INFORMED JUDGMENT. Standard setting necessarily involves judgment and it’s important to include training and practice so participants are making informed judgments. Informed judgment is needed because SPSS, SAS, Excel, etc., cannot set standards. Trained participants who are subject-area experts as well as other stakeholders are in the best position to determine what’s good enough and set the standard.
For the participants to make informed judgments, they need to be very clear on the **purpose** of setting standards, the **process** that they will undertake to set the standard. They need **content knowledge**. They need to know the **consequences**: how will the standards be used, what impact the standards will have on the students and the program. They need **knowledge of the students**. Specifically, they should know

- Characteristics of typical students in the group
- Students’ typical educational experiences
- Students’ typical performance on the task at hand

Some of this information can be conveyed during the orientation and training session.
The second important concept in standard setting is the “borderline student.” The borderline student is used because there needs to be a cut-off point where students above the cut-off score are in one performance category and students below are in a different category.

When you facilitate a standard setting session, you ask the participants to have the borderline student in mind when they set score expectations. These students are just good enough. They may also be called the “just qualified student” or the “minimally competent practitioner” if you are in education or engineering.

The “borderline student” is always difficult for at least one person in the group to understand. A graphic can help.
Here’s an example. If the evaluation instrument, like a test, has a possible score range of 0 to 200, and if the participants are setting one standard—proficient—then they may determine that the borderline student, the just-barely competent student will score 125. Then, 125 becomes the performance expectation, the cutscore. Students scoring 125 or higher are considered proficient. Score of 124 or below, not proficient. The participants think about the students’ educational experiences, the range of performance, the range of knowledge and skill, and then they have in mind the student(s) who are just barely at the proficient level. With the just barely proficient, borderline student in mind, the participants set the standard or cutscore.

My recommendation: when describing the borderline student, use the phrase “will provide a correct response” or “will score on the rubric.” Avoid the phrases “should provide” and “should score” because these typically bring out more personal, emotional reactions.

IMPORTANT: It’s very useful for the facilitators, assessment committee, and any others involved with overseeing session to have a conversation prior to the session regarding the appropriate conception of the hypothetical borderline student group: (a) reflect the participants’ expectations regardless of the status quo or (b) reflect the group of actual students.

Participants will typically ask something like this, “Is the borderline student’s score what I desire the borderline to be? Or is the borderline student’s score what I see given my experiences with seniors?” This question asks whether the participants should envision an
ideal group of students or envision the status quo. My response is similar to this: “Use your experiences with seniors, plus what you’ve learned from others in the training session, your knowledge of students’ educational experiences and learning opportunities, and your knowledge of employer and community expectations of recent graduates, to envision the graduating senior who is borderline meeting expectations in written communication.”
It’s also possible to set multiple standards that give meaning to the scores. In this case, it’s advanced, proficient, and approaches proficiency. The participants will go through three standard-setting processes, one time to determine the score for borderline proficiency students, then again for borderline advanced students, and again for borderline approaches proficiency students.
4. THE SESSION
At the session itself, you’ll need a panel of people who can make informed judgments. They’ll need to understand the standard-setting process, the purpose of standard setting, what the consequences are for the decisions they make, and they need to have subject matter knowledge and knowledge of the students.

A rule of thumb is a panel of 10-15 people with 70% faculty and 30% stakeholders. Stakeholders may include employers or faculty from other degree programs. They may also include students, but care needs to be taken with student panelists regarding test security and with students’ subject matter knowledge. Do they know enough to be able to make informed judgments?
In all standard setting sessions . . .

Orientation. Facilitator explains
1. Purpose
2. Consequences (use of results)
3. Characteristics of the student group, the task, the scoring procedure
4. “Borderline student”
5. Standard-setting process

Example—using student papers for written communication assessment

In program and institutional assessment aimed at using results for program improvement:
1. The purpose is to set performance expectations for seniors. We have collected random samples of seniors’ capstone writing projects and a group of faculty scored them. We know how many students scored 1, 2, 3, and 4 using the Written Communication VALUE rubric. However, we have not decided, as a faculty, what’s good enough. Does a “4” on the rubric mean seniors meet our performance expectations? Or a “3”? “2” “1”? The purpose of today’s session is to figure that out by setting a performance expectation or a “cutscore” on the rubric.
2. The established performance expectation will be used to help guide program decision making. No individual student’s graduation will hinge on a score. Instead, the expectation will let us know the percent of students who exit having met our expectations and the percent who have not. Once we know that, faculty groups can decide what actions are needed: celebrate success, change program policies, advocate for different writing assignments, increase faculty development, etc.
3. Today we’re looking at exit performance expectations for undergraduates in the area of written communication. The students are seniors, near graduation, who have completed all writing requirements. The student work is from the capstone course and students were given feedback on their project papers and were able to revise. The students were also given the rubric so they knew how the writing would be evaluated. A group of faculty members used the rubric to evaluate the student work for program assessment. Our goal in today’s session is to figure out what those rubric scores mean.
Specifically, you’ll help determine the score on the rubric that is required for us to say the student work meets our expectations. The performance expectation will take the form of a score on the rubric.

4. To do this, I will ask you to use your knowledge of seniors, of program writing requirements, and of students’ performance levels, to envision a “borderline student” – a student who is just good enough. You’ll use your understanding of a borderline student to establish the performance expectation.

5. These are the steps we’ll take today:
   1. First we’ll discuss the performance descriptors and the rubric.
   2. Then, I’ll ask you to share your scores on the student work and rationale behind the scores. After discussion and reflection, you will have an opportunity to revise your scores.
   3. Next, you’ll estimate the average score that a borderline student will receive (Round 1)
   4. I’ll show you the group average and the distribution of everyone’s individual scores. I’ll ask for volunteers to explain their average score.
   5. After that discussion, you’ll have an opportunity to revise your score. The final group average score will be the performance expectation for seniors.
   6. Finally, before you leave, I ask that you please complete an evaluation form.
Thank you for listening. Now it’s time for you to get involved and do some work.

I’ll start with the situation in which students in a program are creating written, oral, or other performance products that are evaluated using a rubric.
Rubric-based Method (1/3)

1. Orientation: purpose, consequences, borderline student, etc.
2. Participants practice using the rubric to score student work; share & reflect

Mean estimation (derivative of the Angoff method)

This is the scenario. You are part of a panel of faculty and stakeholders. You’ve finished the orientation in which the purpose, consequences, borderline student and performance descriptions were discussed. The next step is to familiarize yourselves with the scoring process and student work. You will read student work, score, reflect, and revise scores as needed.

[This standard setting technique is called *mean estimation* and is a derivative of the Angoff method.]
Your Turn

HANDOUTS:
Rubric
Student paper

Your turn!
The facilitator explains that the sharing and discussing is done to encourage reflection and thoughtfulness, to better understand one’s own rationale; to add to each others’ understanding.

It is NOT to force consensus.

Rubric-based Method (2/3)

3. Participants estimate the average score that the borderline student will receive on the rubric (to one decimal place)
4. Collect participants’ “average score”
5. Share and discuss
   • Show group average and individual distribution
   • If available, share actual student results to give participants an indication of consequences of particular cut scores

Mean estimation (derivative of the Angoff method)
It’s time to complete the recording form.

Target Student Group & Borderline Student-- things to keep in mind:
What do you know about the group of students?
- Seniors, ready to graduate
- Completed all or nearly all of the degree requirements
- The performance descriptors indicate that student who meet expectations can construct a sustained, coherent written piece.
- Students were able to get feedback and revise their papers; papers contributed significantly course grade so motivation is assumed.

(Most of the participants will have subject-area expertise and can better evaluate correctness of content and they will have had experience with seniors or recent graduates.)

Imagine the borderline student who is just good enough. What average score will the borderline student receive on the rubric? Use up to one decimal place.

Activity Options:
Okay, turn to your table mates and find out why they assigned that number of students to each category.
Now, change your scores if you like, or not.
Discuss. Participants share rationale. Listen, reflect, understand own and others’ perspectives.

Show normative data:
- Their score in relationship to others
- What the group cutscore would mean in terms of actual student results (if available)

After discussion and sharing of normative data, participants revise their score.

See Excel spreadsheet.
6. PAPER SELECTION METHOD

Standard-setting Methods
Another option with constructed response tasks is the paper selection method. In this case, the participants do NOT base their final judgment by examining the rubric. Instead, they look at student work that has already been scored by another group of faculty. The participants do NOT know what scores were given. Their task is to review the student work in light of the outcome and the performance description. They select student work that they view as representing the borderline student.
Then, they discuss their results. They may revise. Then they go home. The facilitators take the student work that the panelists decided was “just good enough” and look up the scores that the student work received. They average the scores to obtain the standard or cutscore.
Next I’m going to talk about how a program can set standards if it has created its own multiple-choice exam.

If a program is using an externally-created exam, the exam creators have gone through this sort of process already and will provide either comparisons between students such as a ranking or percentiles, or provide information about the number of students in each performance category.

But if a program is using its own exam, here’s what you can do.
The orientation is the same as in other standard setting approaches. I highly recommend that the participants take the exam, under the same exam conditions if possible. For a multiple-choice exam, the participants examine each question and consider item difficulty, student characteristics, learning opportunities, etc. They use the concept of the borderline student to assign a probability to each question: “what is the probability that a borderline student will answer the question correctly?” 1.0? .5? .1?

The facilitator leads the participants in practicing assigning probabilities to each question in light of considerations such as item difficulty and student learning opportunities.

After participants are comfortable, they independently estimate the probability that a borderline student will answer each test item correctly. His/her probabilities are summed to arrive at a participant standard or cutoff score (see Excel spreadsheet).
The facilitator leads the discussion about each question and participants are told their average and the group’s average. Other normative data such as how the group’s initial average relates to actual student performance are provided if available. Participants revise their probabilities after discussion and a final average standard or cutscore is calculated.
Things to keep in mind:
What do you know about the group of test-takers?
- Seniors in POLS, ready to graduate
- Completed all or nearly all of the degree requirements
- The performance descriptors indicate that they should be able to apply terms, methods, theories in American, comparative and international politics and global relations.
- Assume the exam conditions were good and students were motivated to do well.

(Most of the participants will have subject-area expertise and can better judge the difficulty of the questions. You’ll have to do your best.)

Imagine the borderline student who is just good enough. What is the probability borderline students will answer correctly? (e.g., .5 or 50% probability).

See Excel spreadsheet.
After the panelists have set the standards, the scores will have meaning. The program will know that a score of “3” or “25” means acceptable performance. The next question will be, have enough students scored in the “proficient” range? Have enough scored in the “advanced” range? I’ll provide some suggestions for establishing benchmarks or targets for success.
The program can set a target for program success. The target usually varies depending on the outcome. For medical doctors, understanding the ethics of the doctor-patient relationship is essential and thus the target is set high, 100%.

A common practice is that in high-stakes decision making or crucial outcomes, err on the side of a higher standard. So the target is 100% for the doctor-patient ethics outcome because this is a crucial outcome. With many program assessment outcomes, the stakes are not as high and 95% or lower may be appropriate.

### Set a Target for Program Success

**Examples**
- 95% will meet expectations (i.e., score “4” or higher)
- 100% will meet expectations on doctor-patient ethics
To set targets with locally-developed instruments or embedded assignments, the program can use its past performance to set targets.
In cases where the program uses a licensure exam that is run by an external organization, the program can set targets based on the national pass rate. Many national exams disaggregate the data; use data that is as close to your program as possible. E.g., use Public-doctoral granting institutions if that is your institution type.

Professional accreditation agencies can have requirements. For example, our Law school has set a 75% pass-rate on the bar which is the accreditation requirement.
If the program uses a standardized exam that provides national comparison data, you can use that. Again, see if you can get national data on institutions that are similar to yours.
See handout. Any questions about these?
What questions do you have?
Mahalo (thank you)!

Monica Stitt-Bergh, PhD
bergh@hawaii.edu

Assessment Office
University of Hawai‘i at Mānoa