Using the Angoff Method to Set Cut Scores

Alan Wheaton, Instructional Systems Specialist
USCG Maritime Law Enforcement Academy
Charleston, SC

Jim Parry, CPT, Test Development Manager
USCG Training Center
Yorktown, VA

Title: Test Defensibility in the US Coast Guard - Using the Angoff Method to Set Cut Scores

Background: In order to distinguish between those who are competent to do the job and those who are not, trainers need to define the acceptable levels of competency as they relate to performance. The Angoff method uses a systematic and documented approach to establish defensible pass/fail scores. Used for over 30 years, this focus-group approach to standard setting has been widely accepted by testing professionals and courts.

The Angoff method is easy to implement and can be perfected by novice users with only minimal training. Attendees will be provided a handout that includes a five step method that will guide them in the design, development and implementation of the Angoff method for their organization. Attendees will leave the workshop with a newfound appreciation of the Angoff method's value in the measurement of true competency.

During this session, the audience will be presented a case study from the Coast Guard Maritime Law Enforcement Academy's (MLEA) Basic Boarding Officer Course. The MLEA will compare their traditional (arbitrary) method of establishing written test cut scores to that of a criteria-referenced technique called the Angoff method. The facilitators will discuss how the USCG is moving their enlisted advancement testing to a more defensible stance by using the Angoff method for establishing cut/pass scores.

Kick-off questions:

- How do you know your test takers are really minimally competent?

- What is a passing score on a test?

- Can you defend your cut or passing score in a court of law?

THE ANGOFF METHOD

Introduction

Development

This section on "The Angoff Method" was developed by the Curriculum Division, Maritime Law Enforcement Academy, Charleston, SC.

References

The information in this section can be found in the following references:

American Society for Training and Development. (2006). Test design and delivery. ACT, Inc.

Angoff, W.H. (1971). Scales, norms and equivalent scores. *Educational Measurements*. Washington, DC: American Council on Education

Ricker, K. (2006). Setting cut-scores: a critical review of the Angoff and modified Angoff methods. *Alberta Journal of Educational Research*, Vol. 52, No. 1, 53-64

The standards for educational and psychological testing. Retrieved December 2011 from:
http://www.apa.org/science/programs/testing/standards.aspx

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement*, Vol. 19, No. 4, 205-308

U.S. Coast Guard. (2008). Performance Qualification Guide, Volume 9

Forms

The forms used during implementation of the Angoff Method are:

- Test Item Rating Form
- Expert Ratings Spreadsheet

Templates for these forms are located at the end of this section.

Overview of the Angoff Method

Introduction   To be legally defensible and meet the Standards for Educational and Psychological Testing, cut scores for tests cannot be arbitrarily determined.

The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) jointly developed the Standards for Educational and Psychological Testing. In addition to providing testing standards, the *Standards* also addresses professional and technical issues of test development and use, and presents measurement trends affecting validity of tests.

Validity / Reliability   For a test to be legally defensible, two standards must be met:

1. Validity: The test must measure what the students are expected to know. This is accomplished by writing test questions that align with the objectives.

2. Reliability: The test must produce consistent results time after time. That is, the test should produce the same score if administered to the same students again and again.

William H. Angoff   William H. Angoff was a research scientist and author who lectured on measurements used in testing and scoring. In 1971, Angoff co-authored the *Educational Measurements* book, where he wrote in a footnote:

"…keeping the hypothetical 'minimally acceptable person' in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration."

"… ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly."

This footnote was the origin of the Angoff Method, a standard-setting process designed to support the defensibility of a cut score.

Applying the Angoff Method

The Angoff Method

The Angoff Method is a process that determines how often a minimally qualified performer would answer a test item correctly. A panel of experts is chosen to review test items and estimate the probability that a minimally qualified performer would answer the items correctly. The estimates for each test item are averaged, and those averages are used to determine the cut score.

While the Angoff Method can be used for performance tests, the information provided in this section applies to written/computer-delivered criterion-referenced assessments only.

Reviewing these types of assessments using the Angoff Method is a dedicated project. Raters must be chosen, a site must be available, and time must be afforded. There are five steps involved:

1. Select the raters.
2. Take the assessment.
3. Rate the items.
4. Review the ratings.
5. Determine the cut score.

Step #1: Select the Raters

Select at least five (5) Subject Matter Expert (SME) raters, and gather them at a common location where they can work both independently and together. Ideally, 10 or more raters are encouraged, but logistics and availability may impede that number.

To select raters, choose SMEs with the following proficiencies:

- Familiarization with the tasks the test will assess
- Knowledge of the skill sets of persons who will perform those tasks
- Ability to pass the existing test at the current cut score
- Ability to edit test items for clarity, accuracy, spelling, and grammar

Additionally…

- For Rating Advancement Tests (RAT), choose SMEs who are at least one grade higher than the examinees (i.e., for E-5 and E-6 RATs, use E-7, E-8, and/or E-9 SMEs).
- For "C" school assessments, merchant mariners examinations, etc. the judges should be selected from a pool of successful course graduates, instructors or licensed merchant mariners who are considered experts on the subjects in the assessment.
- Try to assemble a diverse group of SMEs (e.g., different races, genders, ages, and educational backgrounds).

Applying the Angoff Method

**Step #1:
Select the
Raters (cont.)**

Ideally, the number of raters should be at least 10 to keep the variance among ratings low; the more raters involved, the more accurate the cut score will be.

All judges/raters are required to execute a non-disclosure statement that will become part of their official personnel record to discourage possible compromise of assessment items. The standard non-disclosure statement is included as an appendix to this SOP.

<u>**Note**</u>: The use of Accomplished Performers (APs) is discouraged, as APs may have difficulty relating to the concept of "minimally qualified performer."

**Step #2:
Take the test**

Have the raters take the test using the current cut score, if one has been established. Obtain feedback from raters on objectives, wording, and design of test items. If items need to be revised, do so before the rating process begins.

**Step #3:
Rate the Items**

Prior to beginning the actual rating process, conduct an orientation:

- Provide the definition of a "minimally qualified performer."

- Provide instructions on how to rate the test items.

- Explain the rating process.

Minimally Qualified Performer

A minimally qualified performer is:

- One who performs the task in the field; **not** a student

- One who has the least amount of education and experience necessary to perform the task

- One who meets standards, though barely

- One whose task performance is borderline, but acceptable

In addition to the criteria listed above, factors specific to the job/tasks may be introduced to further identify a minimally qualified performer.

Rating Test Items

Instructions for rating test items are as follows:

- Review test items individually.

- Estimate the number of minimally qualified performers out of 100 who would answer the question correctly.

Applying the Angoff Method

**Step #3:
Rate the Items
(cont.)**

- Record that number on the Test Item Rating form (figure 1) under the "Percentage Correct" (i.e., if 80 out of 100 minimally qualified performers would answer a question correctly, then the percentage correct is 80%).

TEST ITEM RATING			
COURSE NAME:		**TEST NAME**:	
RATER NAME:		**DATE**:	
Instructions: Review each test item. Determine the probability that minimally qualified performers would answer the item correctly. Do not rate higher than 95% nor lower than 25%. Use increments of five (e.g., 80, 65). **Minimally Qualified Performer**: One who has the least amount of education and experience necessary to perform the task.			
TEST ITEM #	**PERCENTAGE (%) CORRECT**	**TEST ITEM #**	**PERCENTAGE (%) CORRECT**

Figure 1. Test Item Rating Form Example

- Estimates should not be higher than 95 nor lower than 25. Not even strong performers are expected to earn a perfect score of 100; and minimally qualified performers can correctly guess an answer 25 percent of the time (for a test item with four distracters).

- During the process, raters should periodically review the concept of a minimally qualified performer to ensure estimates are as accurate as possible.

The Rating Process

Guidelines for implementing the rating process are as follows:

- Give the raters the test items along with the Test Item Rating form.

- Do **not** provide raters with the answer key. This could unduly influence the raters by causing them to underestimate the item difficulty.

- Separate the raters and have them provide estimates for each test item. Allow approximately two (2) hours for a 100-item test.

- Reconvene raters and proceed to the next step.

Applying the Angoff Method

**Step #4:
Review the
Ratings**

Collect the raters' Test Item Rating forms and enter the results in the Microsoft® Excel Expert Ratings Spreadsheet (figure 2):

- Enter the percentages for each test item under the respective rater's name.

- Tabulate the average percentage correct for each test item by adding the raters' percentages and dividing by the number of experts.

- Determine the standard deviation; click in the cell, and from the Formulas tab, select More Functions/Statistical/STDEV.

Test Item	Average Percentage Correct	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name		Standard Deviation
1	76	90	80	50	70	90		16.73320053
2	68	65	70	60	75	70		5.700877125
3	74	85	85	60	60	80		12.94217911
4	85	85	90	80	85	85		3.535533906
5	79	95	85	60	80	75		12.94217911

Figure 2. Expert Ratings Spreadsheet Example

Different estimates from raters for the same test item are to be expected. Arbitrariness can result from diverse conceptions of mastery of the task, various interpretations of the learning objectives, misunderstanding of the test item etc.

Standard deviation reflects the amount of agreement/disagreement among the raters for each test item. A low standard deviation indicates a high agreement among raters (see item #4 in Figure 2 above). A high standard deviation (see item #1 above) is grounds for further examination of that test item.

For any test item whose standard deviation exceeds 10, raters should discuss the reasons for variations in the estimates. The intent of the discussion is to increase agreement among the raters. By discussing how the raters arrived at such different conclusions for a test item, they might decide to re-evaluate their estimates.

Re-Evaluate Test Items

After discussion, separate the raters and have them rate any test items with standard deviations above 10. Collect the ratings and enter them on the spreadsheet.

Applying the Angoff Method

**Step #5:
Determine the
Cut Score**

Once test items have been re-evaluated and the estimates have been entered into the Expert Ratings Spreadsheet, review the sheet for the following:

- If a rater provided the same rating for every test item, consider eliminating those ratings.

- If a rater continually provided ratings that were very dissimilar from the other raters, consider eliminating those ratings.

- If an outlying standard deviation for a test item remains, consider another discussion/re-evaluation session.

  <u>Note</u>: Even if disagreement persists, the average percentage for that test item can be factored into the cut score.

Calculate the cut score by adding the numbers in the "average percentage correct" column and dividing by the number of test items.

In Figure 3, there were five raters evaluating five test items resulting in an average score of 76.4 percent. Round that figure to determine the **cut score**: 76 percent.

<u>Note</u>: When rounding a figure, round down for 0.1 to 0.4; round up for 0.5 and above.

Test Item	Average Percentage Correct	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name		Standard Deviation
1	76	90	80	50	70	90		16.73320053
2	68	65	70	60	75	70		5.700877125
3	74	85	85	60	60	80		12.94217911
4	85	85	90	80	85	85		3.535533906
5	79	95	85	60	80	75		12.94217911
		↓	↓	↓	↓	↓		
Average	**76.4**	84	82	62	74	80		

Figure 3. Calculating the Cut Score

Conclusion

Test item challenges may prompt the need for revision. Any time a test item must be revised, for whatever reason, the Angoff Method should be applied.

TEST ITEM RATING

COURSE NAME:		TEST NAME:	
RATER NAME:		DATE:	

Instructions: Review each test item. Determine the probability that minimally qualified performers would answer the item correctly. Do not rate higher than 95% nor lower than 25%. Use increments of five (e.g., 80, 65).

Minimally Qualified Performer: One with the least amount of education and experience necessary to perform the task.

TEST ITEM #	PERCENTAGE (%) CORRECT	TEST ITEM #	PERCENTAGE (%) CORRECT

EXPERT RATING SPREADSHEET

COURSE NAME:				TEST NAME:				
ANGOFF FACILITATOR NAME:				DATE:				

Test Item #	Average Percentage Correct	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name		Standard Deviation
Average:								

Question*mark* Perception Test

Rater Exercise

Estimate the number of minimally qualified performers out of 100 who would answer the question correctly. Use increments of 5 - lowest number is 25, highest number is 95.

Angoff Score	Test Item

	1. A test item that may have more than one correct response is known as a _____ item.
	   a. multiple choice
	   b. likert scale
	   c. survey matrix
	   d. multiple response

	2. The tool provided by Question*mark* Corporation to license holders who subscribe to a support plan that allows remote entry of test items and assessments without the need to load software locally is Question*mark* _____.
	   a. Connectors
	   b. Live
	   c. To Go
	   d. OnDemand

	3. Question*mark* _____ provide a great introduction to the company's technologies and services.
	   a. Breakfast Briefings
	   b. White Papers
	   c. Mobile Apps
	   d. Connectors

	4. The simplest way to enter test items into Question*mark* Perception Windows Authoring is by using _____.
	   a. item wizard
	   b. content manager
	   c. question wizard
	   d. advanced editor

	5. Which report in Question*mark* enterprise reporter provides a comparison of the differences in results between groups of participants?
	   a. Gap Report
	   b. Grade Book Report
	   c. Coaching Report
	   d. Transcript Report

6. Which of the following is NOT a delivery option for Question*mark* assessments?
    a. Schedule for web delivery
    b. Schedule for delivery at test center
    c. Schedule for Questionmark to Go
    d. Schedule for e-mail delivery

7. Which question type provided in Question*mark* Perception allows the participant to place a single graphical marker on a single image to indicate the correct answer?
    a. Point and click
    b. Hotspot
    c. Drag and drop
    d. Click and drag

8. The method provided by Question*mark* Perception to enable the author to classify questions by specific criteria such as difficulty, metric, acceptability, etc. is known as a/an _____.
    a. IRT
    b. metatag
    c. QID
    d. item tag

9. Topics, assessments and assessment folders can be exported from a repository to external files called _____.
    a. Qpacks
    b. SCOs
    c. Qfiles
    d. QMLs

10. Within Question*mark* Perception, graphic, video or sound files used to provide stimulus to the participant when answering questions are known as _____.
    a. content
    b. resources
    c. outcomes
    d. feedback

What's the score?
Did you really pass?

Using The Angoff Method to Set Cut Scores

Alan Wheaton, Instructional Systems Specialist
USCG Maritime Law Enforcement Academy
Charleston, SC

Jim Parry, CPT, Test Development Manager
USCG Training Center
Yorktown, VA

2012 Users Conference
New Orleans ⚜ March 20 - 23

Introductions!

- Civilian
- Military
- Educators
- Teachers
- Administrators
- Instructors
- Designers
- Other

"Kick-off questions"

1. How do you know your test takers are really minimally competent?

2. What is a passing score on a test?

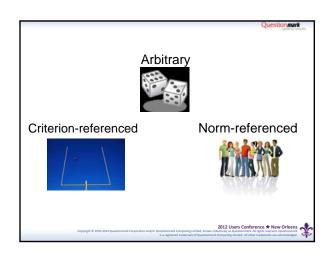3. Can you defend your cut score in a court of law?

Slide 1

"Agenda"

- Discuss types of cut scores

- Discuss the Angoff method

- Present case study

2012 Users Conference ★ New Orleans

Slide 2

Who has experience establishing cut scores?

What methods were used?

2012 Users Conference ★ New Orleans

Slide 3

Arbitrary

Criterion-referenced   Norm-referenced

2012 Users Conference ★ New Orleans

Who is William Angoff?

- Distinguished research scientist at the Educational Testing Service (ETS) for over 40 years
- Harvard, Purdue, U.S. Air Force
- Scholastic Aptitude Test
- Prominent contributor to educational measurement

1919-1993

Why use the Angoff method?

- Defensible
- Easy
- Minimal training
- Widely accepted

Five step (modified)Angoff method:

1. Select the raters.
2. Take the assessment.
3. Rate the items.
4. Review the ratings.
5. Determine the cut score.

Important!

Prior to applying the Angoff method be sure the test is valid and reliable.

Step 1: Select the raters (SMEs)

- **5-12 Subject Matter Experts (SMEs).**
- **Diverse group (geographic location, age, gender, race, etc.).**
- **This step is critical to the success of the process.**

Step 2: Take the assessment

- **SMEs take the test.**
- **The average score can be used to ensure the final passing score is not set higher than the average score obtained by the SMEs.**
- **This step increases the defensibility of the final score.**

Step 3: Rate the items

- **Train the SMEs.**
- **Describe the characteristics of minimally competent performers:**
  - o **minimally qualified employee on the job**
  - o **not a student**
- **Describe how to rate items:**
  - o **(increments of 5, low 25, high 95)**

Handout page 4 & 5

Step 4: Review the ratings

- **Collect the SMEs test item rating forms**
- **Enter results into a spreadsheet**
- **Determine standard deviation**
- ***Re-evaluate* test items that exceed standard deviation (10)**
- **Enter revised results to the spreadsheet**

Handout page 6

Step 5: Determine the cut score

- **Add the numbers in the "test item" column and divide by the number of SMEs. Enter each test item average score in the "average percentage correct" cell.**

- **Calculate the cut score by adding the numbers in the "average percentage correct" column and dividing by the number of test items.**
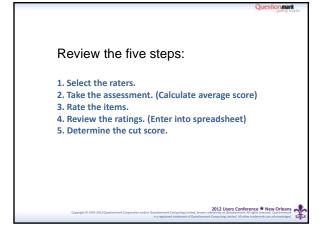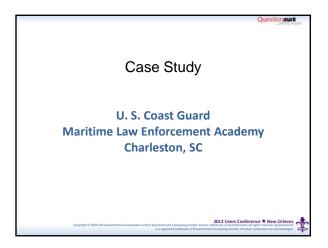
Handout page 7

Let's try it out!

- Take the Question*Mark* test at the end of the handout. (Calculate average score)
- Rate the items.
- Review the ratings. (Enter into spreadsheet)
- Determine the cut score.

Review the five steps:

1. Select the raters.
2. Take the assessment. (Calculate average score)
3. Rate the items.
4. Review the ratings. (Enter into spreadsheet)
5. Determine the cut score.

Questions?

Case Study

U. S. Coast Guard
Maritime Law Enforcement Academy
Charleston, SC

Background

MLE Academy Written Test

2008 - Cut score was arbitrarily set at 80%

2009 - Applied for accreditation
Federal Law Enforcement Training Accreditation (FLETA)

2009 - Angoff: raised cut score to 85%

Data

Prior to Angoff - 4% failure rate

After Angoff - 12% failure rate (first academic drop out)

After Angoff – of the 12% that failed,
76% of the failures scored between 81 and 85%

Results

Had we not raised the cut score from 80% to 85%, 110 of our students (over an 18 month period) would have passed the exam at a less than minimally qualified standard.

Angoff method was considered a "best practice" by the Federal Law Enforcement Training Accreditation (FLETA) Board.

MLEA Future Plans

Continue analyzing Level 2 data and apply the Angoff method to appropriate exams.

Explore the use of the Angoff method for performance based evaluations.

Mr. James Parry
U. S. Coast Guard

Enlisted Advancement Testing

Questions?