Setting Fair, Defensible
Cut (Passing) Scores

James R. Parry, M.Ed. CPT
Owner/Chief Executive Manager
Compass Consultants, LLC
Acknowledgments

• Photos and some illustrations used fall within the creative commons license
• Stick figures courtesy of E-Learning Heroes
Cut Scores

Who says I failed?
Cut Scores

• Assist in interpretation
• Sometimes called the passing score can be used for either raw or scaled scores
  • May be used to determine pass/fail (CRT)
  • May be used for decision for placement (NRT)
Establishing Cut Scores

• When cut scores are used they should be set as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force

• To be legally defensible and meet the Standards for Educational and Psychological Testing, a cut score cannot be arbitrarily determined, it must be empirically justified

• No single method can be used to determine cut score for all tests
Establishing Cut Scores

• There is no single set of rules to establish defensibility
• Cut scores may be set as high or as low as needed to meet organizational requirements
• Establishing cut scores involve professional judgments as well as technical and empirical considerations
• Should use a sufficiently large and representative group of judges to ensure validity
• Procedure used must be documented
Minimal Acceptable Competence (MAC) Level

- The level of performance on the test indicative of minimal competence
  - Bare minimum – the bottom of the qualified barrel
  - This is NOT the best or most qualified
Various Methods

No one is “most” correct!
Recognized Methods

• Angoff/Modified Angoff
  • Most widely used

• Nedelsky

• Ebel

• Jaeger

• Bookmark
Description of Methods

• Nedelsky Method
  • Obvious distractors eliminated
  • Results in a modified chance score for each item

• Ebel Method
  • Items categorized by relevancy and difficulty
  • Item categories evaluated based on probable MAC candidate performance
Description of Methods

• Jaeger Method
  • SME judgments combined with actual SME performance data

• Bookmark Method
  • SME judgments combined with actual candidate test performance data
    • Data rank-ordered high to low
    • Select items that a MAC should get correct
Description of Methods

• Angoff Method
  • Item performance probability determined by panel of expert judges
  • Item probabilities summed

• Modified Angoff Method
  • Item necessity and difficulty levels determined
  • Item performance probability determined
  • Results calculated
  • Combination of Angoff and Ebel methods
What is a Cut (Pass) Score?

• Based on MAC level
• Based on rationale of expected test performance
• Designed to result in a cut or pass point that represents the threshold between those candidates who can do the job and those who cannot
• Cut/Pass score selected must be defensible as the most/best rational that a reasonable person on the job would accept
Angoff Method

• Most widely used and accepted
  • 40+ years of research
• Item performance probability determined
  • Minimally competent performer
• Item probabilities summed
  • Sum \( \div \) number of evaluators = cut score
Modified Angoff Method

5-step method
The Modified Angoff Method is a way to set fair cut or passing scores on tests. It has been used for well over 40 years and has been accepted as a best practice in legal test defensibility situations.
The Modified Angoff Method works with the assumption that we have reasonable and consistent expectations of acceptable proficiency within the work force.
Based on the judgment of a panel of experts, this method is designed to result in a cut or pass point that represents the threshold between those candidates who can do the job and those who cannot.
There are basically 5 steps to the Modified Angoff Method.
STEP 1

• Select raters familiar with competencies covered by test and with performance level for masters of these competencies
  • 3 - 5 is adequate
  • 9 -12 maximum
  • Diverse group (geographic location, age, gender, race, etc.)
You are considered experts for this subject. You will be using the Modified Angoff method to set a fair cut score for the final exam.
STEP 1 (cont.)
• Conduct face-to-face meeting of judges to set the cut score
  • If face-to-face is not possible a virtual meeting should be conducted
  • Judges come to consensus regarding definition of “minimally competent performer”
    • Not “incompetent”
You have to decide what you consider a minimally competent performer to be. Basically – what level of performance would you be happy with.
Can you give us an example of how to decide what that is?
Good Question Hunter! This is the most important and probably most difficult step in the process.
Think of a new, just qualified, welder, first day on the job.
Let’s think of four welders...

- NOT Competent
- Minimally Competent
- Competent
- Highly Competent
When asked to strike an arc, a welder who is NOT COMPETENT may respond with, “Strike a what?”
When asked to strike an arc a **MINIMALLY COMPETENT** welder may respond with, “No problem, but could you stand next to me in case I have trouble maintaining the arc?”

I’d be happy with an apprentice who asked me that. They have good initiative.
When asked to strike and arc a COMPETENT welder may respond with, “No problem! Would you like me to continue running the bead?”

Sounds like most of the welders I know!
A HIGHLY COMPETENT welder may respond with, “No Problem! Would you like me to train the apprentices while I’m showing you?”
STEP 2
• Each judge is given copy of all questions
  • Each should complete “exam” as test taker would
    • Verifies validity
    • Cut score not set higher than average raters score!
  • Correct response should be shown after completion
Part of the rating process requires that you each take the test as a real test taker would. Please read each item carefully and answer based on your knowledge. Don’t worry, your score won’t be released to anyone.

Whew! I was worried about losing my job if I failed!
You didn’t do too bad,  
Your average score as a group was 85%
STEP 3
• Estimation process
  • Probability estimate can never be less than .25 (25%) if there are 4 choices for a multiple choice question
    • This is minimum value due to chance guess
    • A 3-response item would have a .33 minimum value
    • T/F & Y/N would be .50 minimum
The next thing you will do is rate each test item. Think about 100 test takers. How many minimally competent performers should get the item correct?
Remember, the MIMIMALLY COMPETENT performer is taking the test, NOT YOU. How many of them would be expected to answer the item correctly?

Boy, this is tough to get in my head!
These are all 4-alternative multiple choice questions so you can’t rate below 25% or above 95%.

25% is the floor because of guessing and 95% is the ceiling because if the MINIMALLY COMPETENT performer would always get it correct it is probably something that should NOT be tested.
So basically, 25% means that probably none of the minimally competent performers would get it right, correct?
Very good observation Ryan! If you consider that 25% is the score achieved by pure chance or guessing that is correct!
Wait a minute! How can we expect a
minimally competent performer to ever get a
95% when our average on the test was only
85% and we are the experts?
Great question Eric! I was hoping someone would catch that. The reason you all took the test was to set the ceiling score which can’t be higher than the average score of the experts. You are right, we can’t expect minimally competent performers to achieve a 95% if the best the experts can do is 85%. So our ceiling is set at 85%.
Technically, any value between 25% - 95% is OK but let’s set some fixed values:

25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95

In this case – the highest we will go is 85
You will work independently and not discuss the test items with each other.
Step 4
- After all items are rated and recorded if any vary among judges by more than Standard Deviation (SD) of 10 they should be discussed
- Weights can be changed as a result of the discussion or original estimates can be retained
I will enter your results into a spreadsheet. We will take a look at the ones that have a standard deviation of more than 10.
<table>
<thead>
<tr>
<th>Test Item QID</th>
<th>Enter &quot; if New, R if Retired</th>
<th>Difficulty Metadata</th>
<th>Average Percentage Correct (Angoff Rating)</th>
<th>Expert 1</th>
<th>Expert 2</th>
<th>Expert 3</th>
<th>Expert 4</th>
<th>Expert 5</th>
<th>Expert 6 Name</th>
<th>Expert 7 Name</th>
<th>Expert 8 Name</th>
<th>Standard Deviation</th>
<th>Unit Cut Score</th>
<th>Moderate Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1.1.1/1</td>
<td></td>
<td>Moderate</td>
<td>72.00</td>
<td>70</td>
<td>80</td>
<td>60</td>
<td>75</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>7.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.1/2</td>
<td></td>
<td>Moderate</td>
<td>72.00</td>
<td>70</td>
<td>75</td>
<td>55</td>
<td>80</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td>10.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.1/3</td>
<td></td>
<td>Easy</td>
<td>79.00</td>
<td>85</td>
<td>80</td>
<td>70</td>
<td>90</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>8.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.1/4</td>
<td></td>
<td>Moderate</td>
<td>60.00</td>
<td>70</td>
<td>60</td>
<td>40</td>
<td>60</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>12.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/1</td>
<td></td>
<td>Hard</td>
<td>69.00</td>
<td>55</td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td>6.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/2</td>
<td></td>
<td>Moderate</td>
<td>53.00</td>
<td>60</td>
<td>50</td>
<td>45</td>
<td>50</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>6.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/3</td>
<td></td>
<td>Moderate</td>
<td>52.00</td>
<td>60</td>
<td>50</td>
<td>35</td>
<td>55</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>10.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/4</td>
<td></td>
<td>Moderate</td>
<td>63.00</td>
<td>75</td>
<td>60</td>
<td>45</td>
<td>60</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>12.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/6</td>
<td></td>
<td>Moderate</td>
<td>73.00</td>
<td>85</td>
<td>70</td>
<td>60</td>
<td>75</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>9.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/7</td>
<td></td>
<td>Moderate</td>
<td>74.00</td>
<td>85</td>
<td>75</td>
<td>70</td>
<td>80</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>9.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/8</td>
<td></td>
<td>Moderate</td>
<td>74.00</td>
<td>85</td>
<td>75</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>6.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.2/9</td>
<td></td>
<td>Moderate</td>
<td>68.00</td>
<td>75</td>
<td>75</td>
<td>70</td>
<td>60</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>7.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.3/1</td>
<td></td>
<td>Moderate</td>
<td>61.00</td>
<td>75</td>
<td>75</td>
<td>55</td>
<td>60</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td>10.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.3/2</td>
<td></td>
<td>Moderate</td>
<td>65.00</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>70</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>8.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.3/3</td>
<td></td>
<td>Moderate</td>
<td>62.00</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>60</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>7.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.3/4</td>
<td></td>
<td>Moderate</td>
<td>62.00</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>60</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>5.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1.3/6</td>
<td></td>
<td>Moderate</td>
<td>74.00</td>
<td>85</td>
<td>60</td>
<td>80</td>
<td>75</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>9.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Standard Deviation**
A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.
What significance does a Standard Deviation of 10 have?

Great question Melissa! To put it simply, the greater the Standard Deviation, the greater the spread or difference in the list of numbers.
Think of *Olympics* or *Dancing With the Stars* Judges. If 5 out of 6 judges score close and the other one is way different - something may be amiss. In this sample the SD is 13.57.
So what happens when the SD is greater than 10?

I was going to ask the same thing!

I see you are thinking Heather! We discuss it as a group to see why there are outliers and either come to a consensus or not.
I noticed the items are being labeled easy, moderate and hard. Where does that come from?

Well Madison, based on an Angoff rating floor of .25 and a ceiling of .95, we divide the score into thirds. This allows us to estimate the difficulty of each item.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Angoff Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>75 - 95</td>
</tr>
<tr>
<td>Moderate</td>
<td>50 – 74.99</td>
</tr>
<tr>
<td>Hard</td>
<td>25 – 49.99</td>
</tr>
</tbody>
</table>
Step 5
• Determine the final cut score
So it looks like we agree that the cut score for this bank of test items will be set at 65%
<table>
<thead>
<tr>
<th>Test Item QID</th>
<th>Difficulty Metag</th>
<th>Average Percentage Correct (Angoff Rating)</th>
<th>Expert 1</th>
<th>Expert 2</th>
<th>Expert 3</th>
<th>Expert 4</th>
<th>Expert 5</th>
<th>Expert 6 Name</th>
<th>Expert 7 Name</th>
<th>Expert 8 Name</th>
<th>Standard Deviation</th>
<th>Unit Cut Score</th>
<th>Difficulty Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1.1/1/1</td>
<td>Moderate</td>
<td>72.00</td>
<td>70</td>
<td>80</td>
<td>60</td>
<td>75</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>7.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/1/2</td>
<td>Moderate</td>
<td>72.00</td>
<td>70</td>
<td>75</td>
<td>55</td>
<td>80</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td>10.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/1/3</td>
<td>Easy</td>
<td>79.00</td>
<td>85</td>
<td>80</td>
<td>70</td>
<td>90</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>8.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/1/4</td>
<td>Moderate</td>
<td>60.00</td>
<td>70</td>
<td>60</td>
<td>40</td>
<td>60</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>12.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/1</td>
<td>Hard</td>
<td>49.00</td>
<td>55</td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td>6.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/2</td>
<td>Moderate</td>
<td>53.00</td>
<td>60</td>
<td>50</td>
<td>45</td>
<td>50</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>6.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/3</td>
<td>Moderate</td>
<td>52.00</td>
<td>60</td>
<td>50</td>
<td>35</td>
<td>55</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>10.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/4</td>
<td>Moderate</td>
<td>63.00</td>
<td>75</td>
<td>60</td>
<td>45</td>
<td>60</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>12.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/5</td>
<td>Moderate</td>
<td>73.00</td>
<td>85</td>
<td>70</td>
<td>60</td>
<td>75</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>9.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/6</td>
<td>Moderate</td>
<td>74.00</td>
<td>85</td>
<td>75</td>
<td>70</td>
<td>80</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>9.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/7</td>
<td>Moderate</td>
<td>74.00</td>
<td>85</td>
<td>75</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>6.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/8</td>
<td>Moderate</td>
<td>68.00</td>
<td>75</td>
<td>75</td>
<td>70</td>
<td>60</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>7.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/2/9</td>
<td>Moderate</td>
<td>64.00</td>
<td>75</td>
<td>75</td>
<td>55</td>
<td>60</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td>10.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/3/1</td>
<td>Moderate</td>
<td>61.00</td>
<td>70</td>
<td>60</td>
<td>55</td>
<td>50</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>8.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/3/2</td>
<td>Moderate</td>
<td>65.00</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>70</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td>7.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/3/3</td>
<td>Moderate</td>
<td>62.00</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>60</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>5.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/3/4</td>
<td>Moderate</td>
<td>62.00</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>60</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>5.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1.1/3/5</td>
<td>Moderate</td>
<td>74.00</td>
<td>85</td>
<td>60</td>
<td>80</td>
<td>75</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td>9.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Standard Deviation**

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

**Unit Cut Score**

65.00

**Difficulty Rating**

25 - 49.99 Hard
50 - 74.99 Moderate
75 - 95 Easy

**In this unit**

20% Easy
66% Moderate
13% Hard
100% TOTAL
By using this method we can ensure that the score achieved by each test taker generally represents their level of knowledge of their assigned job. Considering this assumption we can predict actual expected job performance.
I understand how we are going to set the cut score but isn’t that only going to be valid if all of the test items in the database are used on a single test?
Outstanding observation Sandra! You are absolutely correct! We have developed a spreadsheet that will assist in the design of a test of any length that will maintain the same difficulty and approximate cut score using a pseudo-random item selection technique.
Each item is tagged in the database with hard, moderate, or easy. We then design the pseudo-randomized test based on the outcome of the final test design worksheet.
The spreadsheet compiles the results from each unit in the database and tells us how many items at what level of difficulty to pick to create a fair test every time.
## Final Randomized Test Design Blueprint for: SAMPLE

<table>
<thead>
<tr>
<th>Unit</th>
<th>Unit Cut Score &amp; Difficulty</th>
<th>Items in Unit</th>
<th>% of Total Items</th>
<th>Avail Hard</th>
<th>% From Unit</th>
<th>Avail Mod</th>
<th>% From Unit</th>
<th>Avail Easy</th>
<th>% From Unit</th>
<th>Total # Needed From Unit</th>
<th>Use Hard (Calculated)</th>
<th>Use Hard (Actual)</th>
<th>Use Mod (Calculated)</th>
<th>Use Mod (Actual)</th>
<th>Use Easy (Calculated)</th>
<th>Use Easy (Actual)</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
<td>65</td>
<td>50.00%</td>
<td>5</td>
<td>25%</td>
<td>10</td>
<td>50%</td>
<td>5</td>
<td>25%</td>
<td>15.00</td>
<td>3.75</td>
<td>4</td>
<td>7.50</td>
<td>7</td>
<td>3.75</td>
<td>4</td>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td>66</td>
<td>50.00%</td>
<td>5</td>
<td>25%</td>
<td>8</td>
<td>40%</td>
<td>7</td>
<td>35%</td>
<td>15.00</td>
<td>3.75</td>
<td>4</td>
<td>6.00</td>
<td>6</td>
<td>5.25</td>
<td>5</td>
<td>2.0</td>
</tr>
<tr>
<td>3.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>3.0</td>
</tr>
<tr>
<td>4.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>5.0</td>
</tr>
<tr>
<td>6.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>6.0</td>
</tr>
<tr>
<td>7.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>7.0</td>
</tr>
<tr>
<td>8.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>8.0</td>
</tr>
<tr>
<td>9.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>9.0</td>
</tr>
<tr>
<td>10.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>10.0</td>
</tr>
<tr>
<td>11.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>11.0</td>
</tr>
<tr>
<td>12.0</td>
<td></td>
<td></td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>12.0</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>40</td>
<td>100.00%</td>
<td>10</td>
<td>25%</td>
<td>18</td>
<td>40%</td>
<td>12</td>
<td>35%</td>
<td>30.00</td>
<td>7.50</td>
<td>8</td>
<td>13.50</td>
<td>13</td>
<td>9.00</td>
<td>9</td>
<td>12.0</td>
</tr>
</tbody>
</table>

**NOTE:** If a value appears in the “Total # Needed From Unit” block - you do not have sufficient items in the unit indicated to design a fair test.

### Test Difficulty
- **Test Cut Score:** 66.00
- **Set Desired Test Size:** 30
- **Checksum:** 30

After all Angoff data has been entered on Unit worksheets set the desired test size in the block to the left. Based upon the number of available items, the quantity of Hard, Moderate and Easy from each unit will populate automatically. Use these results to design the test in Questionmark using established Difficulty Metatags. Note: Due to rounding errors in Excel, the unit/Item difficulty totals may require you to round up or down manually to achieve desired test size. Set the actual number desired based upon the calculated results in the columns labeled “Actual” above. The Checksum to the left will alert you if the selected value does not match the desired test size.
So to review the 5-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
An interesting statement was published by the Educational Testing Service (ETS) in their publication: *A primer on Setting Cut Scores on Tests of Educational Achievement:*

“*It is impossible to prove that a cut score is correct.*”

Any comments about that?
Combining Methods

• Use Modified Angoff initially
  • No test result history
• Confirm Angoff performance of each item after sufficient data gathered
• Utilize test performance data to update cut score using Bookmark method
Remember – calculated cut score may be modified by HR requirements and set higher or lower to meet needs!
Why is it important to set a cut score using a recognized method?
Defensibility
Questions?
Thank You!

Website: www.gocompassconsultants.com

E-mail: james.parry@gocompassconsultants.com