Setting Fair, Defensible Cut Scores – More Than One Method to Use

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
Legal Disclaimer

• The presentation may include information about legal issues and legal developments. Such materials are for informational and/or educational purposes only and may not reflect the most current legal developments. These informational/educational materials are not intended, and should not be taken, as legal advice on any particular set of facts or circumstances. You should contact an attorney for advice on specific legal problems or questions.

• Information is provided "as is" without any express or implied warranty of any kind including warranties of merchantability, noninfringement of intellectual property, or fitness for any particular purpose. In no event shall Compass Consultants, LLC., or its agents, officers or attorneys be liable for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information) arising out of the use of or inability to use the information, even if Compass Consultants, LLC has been advised of the possibility of such damage.
References


Expected Passing Score

• What does “expected passing score” actually mean and how is it determined?

• Let’s take for example the “expected passing score” in the antiquated universal scoring system
  • Why is 60% and above a passing score?
  • Is it okay for a student to only retain 60% of every bit of knowledge?
  • Why not 100%, 82%, 76%, or 45%, etc.?
  • Who decides this and how is it determined?
  • Is there a better way to set a passing or cut score that will more accurately reflect actual knowledge retention?
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)
What is a Defensible 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
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
Recognized Methods

Methods based on judgements about test questions (test-centered)
• Nedelsky Method
• Ebel Method
• Jaeger Method
• BookmarkMethod
• Angoff/Modified Angoff Method

Methods based on judgments about a group of test-takers (test-taker centered)
• Borderline-Group Method
• Contrasting-Groups Method
• Up-and-Down Method
• Body of Work Method
Test-Centered Methods

It’s all about d-base (the database, that is...), not people...
Nedelsky Method

• Suited best for multiple-choice and other selected response tests
• Obvious distractors eliminated
  • Not even a “D” student would select
• Results in a modified chance score for each item
• Study by Livingston & Kastrinos, (1982)
  • Concluded method was difficult to generalize about judges’ consistency
  • Consider two-stage process – judges go two rounds and compare results
Ebel Method

• Suited for small scale performance-based assessments
• Uses “borderline” candidate as MAC

• Five basic steps
  1. Select the judges
  2. Define “borderline” knowledge, skills and abilities
  3. Train the judges in the use of the Ebel Method
  4. Collect judgements
  5. Combine the judgements to determine a passing score

• Two stage process
Ebel Method – Stage 1

1. Each judge classifies test items into groups
   i. Easy
   ii. Medium
   iii. Hard

2. Further subdivide items based upon relevance to final outcome
   i. Essential
   ii. Important
   iii. Acceptable
   iv. Questionable
Ebel Method – Stage 2

1. Each judge makes their own judgement about the expected performance of the “borderline” candidate.
   - **Example:** How many candidates should respond correctly to an essential/easy difficulty test item.

2. Results recorded in a twelve-block matrix for each judge

<table>
<thead>
<tr>
<th>Sample Ebel Method Classification Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difficulty</strong></td>
</tr>
<tr>
<td>Relevance</td>
</tr>
<tr>
<td>Essential</td>
</tr>
<tr>
<td>Important</td>
</tr>
<tr>
<td>Acceptable</td>
</tr>
<tr>
<td>Questionable</td>
</tr>
</tbody>
</table>
Ebel Method – Final Step

- Multiply the judged percentage correct for the first category by the number of questions in that category to get the candidate’s expected score for the first category.
- Repeat for the remaining 11 categories.
- Add the expected scores for the 12 categories to arrive at a cut or passing score for the test.

### Sample Ebel Method Final Calculation

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Percent Correct</th>
<th>Number of items</th>
<th>Expected Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essential</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>95</td>
<td>5</td>
<td>.95x5=4.75</td>
</tr>
<tr>
<td>Medium</td>
<td>85</td>
<td>3</td>
<td>.85x3=2.55</td>
</tr>
<tr>
<td>Hard</td>
<td>80</td>
<td>1</td>
<td>.80x1=0.80</td>
</tr>
<tr>
<td>Important</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>90</td>
<td>3</td>
<td>.90x3=2.70</td>
</tr>
<tr>
<td>Medium</td>
<td>75</td>
<td>3</td>
<td>.75x3=2.25</td>
</tr>
<tr>
<td>Hard</td>
<td>60</td>
<td>2</td>
<td>.60x2=1.20</td>
</tr>
<tr>
<td>Acceptable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>80</td>
<td>1</td>
<td>.80x1=0.80</td>
</tr>
<tr>
<td>Medium</td>
<td>55</td>
<td>2</td>
<td>.55x2=1.10</td>
</tr>
<tr>
<td>Hard</td>
<td>35</td>
<td>2</td>
<td>.35x2=0.70</td>
</tr>
<tr>
<td>Questionable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>50</td>
<td>1</td>
<td>.50x1=0.50</td>
</tr>
<tr>
<td>Medium</td>
<td>-</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td>20</td>
<td>2</td>
<td>.20x2=0.40</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>25</td>
<td>17.75 (raw score)</td>
</tr>
</tbody>
</table>

71% (scaled score)
Jaeger Method

• Uses large group of test-takers
• Sample of score profiles of various groups reported to panel of judges
• Judges may believe some scores are acceptable and others are not
• Calculate the median for groups of scores
• Suggest setting cut score at the lowest median of combined groups
Bookmark Method

• Focuses on actual question difficulty vice expected difficulty
• SME judgements combined with actual candidate performance
• Items placed in a book in order of increasing difficulty as determined by item response theory (IRT) results (each gets its own page)
• Judges given course content map
• Judges discuss how each item relates to course content
• Each judge places a bookmark between the item a MAC would likely get correct and the one they would likely get incorrect
Bookmark Method

• Item difficulties of each bookmark averaged
• Judges discuss results
• Judges may reset their bookmark
• Cut score is set
• This method can be used if more than two classifications desired
  • E.G. Novice, Proficient and Advanced
Angoff

• 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
Angoff Method

• Most widely used and accepted
  • 30+ years of use/research

• Item performance probability determined
  • Minimally competent performer

• Item probabilities summed
  • Sum ÷ number of evaluators = cut score
Angoff Method (cont.)

• 5-Steps
  1. Select and gather the judges (raters)
  2. Identify the “minimally qualified performer”
  3. Rate the items
  4. Review the ratings
  5. Determine the cut score
Modified Angoff Method

5-step method
Step 1

• Select raters’ familiar with competencies covered by test and with performance level for masters of these competencies
  • 5 is adequate
  • 8-10 maximum

• Diverse group (geographic location, age, gender, race, etc.)

• Proficiencies of raters:
  • 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
Step 1 (cont.)

- Conduct face-to-face meeting of judges
  - If face-to-face is not possible a virtual meeting should be conducted
- Judges “take” the test under same conditions as a “real” test-taker would
  - Establishes a “ceiling” score – highest rating each item can be assigned
    - Experts can only achieve this score so MAC can’t be expected to
  - Judges provide feedback on wording and design of test items

Note: In the case of a large test item database it may not be practical for the raters to complete the entire item bank due to time constraints so this step may be omitted and noted in the test plan.
Step 2

• Judges come to consensus regarding definition of “minimally acceptable candidate” (MAC)
  • One who performs the task on the job; **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
Step 3

• Estimation process explained
  • 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
Step 3 (cont.)

• Acceptable estimates are .25 through .95 for 4 response items
  • Any value is OK
    • .40 means 40% of minimally competent test takers should respond correctly
    • .50 Fifty-fifty chance that minimally competent performer responds correctly
    • .70 means 70% of minimally competent test takers should respond correctly
    • 1.0 means 100% - so simple that all should get correct
      • Critical content that all should get correct. Why test?
Step 3 (cont.)

• Establish “allowable” percentages
  • Various philosophies
  • Theoretically range from 0 to 1.00
    • Difficult to use entire range

• Widely acceptable to have “set” ranges
  • .25, .35, .45, .55, .65, .75, .85, .95
    • Nothing below .25
    • Nothing above .95
    • 1.00 means wasted test item – all get correct
Step 3 (cont.)

• Examine each stem (question), correct response and all distractors (alternatives)

• Decide how many distractors a minimally competent performer would surely eliminate
  • This provides a “floor” to rating the item
  • If one distractor on a four-choice item will be eliminated, then the lowest possible rating for this item is .33
Step 3 (cont.)

• If choice between remaining distractors is truly random, estimate the Angoff weight as the chance probability between/among the remaining options
  • **Example**: If two choices remain from four distractors then the chance probability would be .50 or 50%
Step 3 (cont.)

• Estimate the difficulty of each item at the minimally competent test-taker level – NOT your level as a judge

• Do NOT estimate the level of a typical test-taker – think of the minimally competent person day 1 on the job

• Set the standard at which the minimally competent performer should be able to answer
Step 3 (cont.)

• Do NOT discuss your ratings of each item
• Work independently
• Read each stem, correct answer and distractor carefully
• Record your rating between .25 and .95 (25% - 95%) for each item next to the item and on the spreadsheet (if provided)
  • Do not set rating higher than the determined ceiling score
Step 4

- After all items are rated and recorded if any vary among judges by more 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
<table>
<thead>
<tr>
<th>Test Item QnD</th>
<th>Enter* If New, Retired</th>
<th>Difficulty Metatag</th>
<th>Average Percentage Correct (Angoff Rating)</th>
<th>Expert 1 Name</th>
<th>Expert 2 Name</th>
<th>Expert 3 Name</th>
<th>Expert 4 Name</th>
<th>Expert 5 Name</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>1.1:01</td>
<td>Moderate</td>
<td>68.18</td>
<td>70</td>
<td>65</td>
<td>70</td>
<td>75</td>
<td>65</td>
<td>60</td>
<td>70</td>
<td>70</td>
<td>4.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1:02</td>
<td>Moderate</td>
<td>53.75</td>
<td>60</td>
<td>55</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>4.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1:03</td>
<td>Hard</td>
<td>49.38</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
<td>45</td>
<td>60</td>
<td>55</td>
<td>45</td>
<td>6.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1:04</td>
<td>Moderate</td>
<td>72.50</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>70</td>
<td>80</td>
<td>75</td>
<td>75</td>
<td>80</td>
<td>13.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1:05</td>
<td>Easy</td>
<td>85.03</td>
<td>90</td>
<td>85</td>
<td>90</td>
<td>80</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>80</td>
<td>5.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2:01</td>
<td>Moderate</td>
<td>53.75</td>
<td>80</td>
<td>65</td>
<td>90</td>
<td>55</td>
<td>60</td>
<td>40</td>
<td>50</td>
<td>50</td>
<td>7.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2:02</td>
<td>Hard</td>
<td>44.38</td>
<td>40</td>
<td>45</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>50</td>
<td>55</td>
<td>45</td>
<td>9.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2:03</td>
<td>Moderate</td>
<td>88.75</td>
<td>95</td>
<td>85</td>
<td>80</td>
<td>80</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>6.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2:04</td>
<td>Hard</td>
<td>68.13</td>
<td>70</td>
<td>65</td>
<td>70</td>
<td>75</td>
<td>65</td>
<td>50</td>
<td>70</td>
<td>70</td>
<td>4.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2:05</td>
<td>Moderate</td>
<td>53.75</td>
<td>80</td>
<td>55</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>4.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3:01</td>
<td>Moderate</td>
<td>49.38</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
<td>45</td>
<td>60</td>
<td>55</td>
<td>45</td>
<td>6.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3:02</td>
<td>Moderate</td>
<td>70.63</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>80</td>
<td>14.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3:03</td>
<td>Easy</td>
<td>87.50</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>80</td>
<td>85</td>
<td>85</td>
<td>90</td>
<td>5.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3:04</td>
<td>Moderate</td>
<td>53.75</td>
<td>80</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>60</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>7.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3:05</td>
<td>Hard</td>
<td>45.63</td>
<td>50</td>
<td>45</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>55</td>
<td>45</td>
<td>9.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4:01</td>
<td>Easy</td>
<td>95.00</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4:02</td>
<td>Moderate</td>
<td>69.38</td>
<td>75</td>
<td>80</td>
<td>80</td>
<td>55</td>
<td>35</td>
<td>75</td>
<td>75</td>
<td>80</td>
<td>16.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4:03</td>
<td>Easy</td>
<td>87.50</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>80</td>
<td>85</td>
<td>85</td>
<td>90</td>
<td>5.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4:04</td>
<td>Moderate</td>
<td>53.75</td>
<td>60</td>
<td>65</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>40</td>
<td>50</td>
<td>50</td>
<td>7.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4:05</td>
<td>Hard</td>
<td>43.75</td>
<td>50</td>
<td>45</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>55</td>
<td>45</td>
<td>6.94</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unit Cut Score: 65.00

Difficulty Rating:
- 25 - 49.99 Hard
- 50 - 74.99 Moderate
- 75 - 95 Easy

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.

5 Easy in this unit 25%
10 Moderate in this unit 50%
5 Hard in this unit 25%
20 TOTAL 100%
Step 5

• Determine Cut Score/Design Assessment – Parry Method

Parry, J.R. (2017)
<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>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>65</td>
<td>20</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>
</tr>
<tr>
<td>2.0</td>
<td>65</td>
<td>20</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>
</tr>
<tr>
<td>3.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12.0</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>40</td>
<td>10</td>
<td>100.00%</td>
<td>12</td>
<td>12</td>
<td>30</td>
<td>30</td>
<td>10</td>
<td>12</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>
</tr>
</tbody>
</table>

**Note:** If there is a blue cell in the "Total # Needed From Unit" block - you do not have sufficient items in the unit indicated to design a fair test.

---

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.
Review of 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
Precautions for Test-Centered Ratings

• Do NOT compromise the test questions!
  • Do not discuss any content with any unauthorized personnel
  • Do not leave questions unattended material
    • Maintain under lock when unattended

• Note: When rating an entire database used to draw random test items it may not be feasible for raters to “take” exam
Test-Taker Centered Methods

People will be people
Borderline-Group Method

• Passing score should be the score expected of a person with “borderline” skills
  • Not quite ready for full performance but not inadequate
• Judges decide what a “borderline” performer “looks” like
• Passing score set at the median (50th percentile) of the borderline group
• Disadvantage – most test-takers are not “borderline” performers
  • Judges may have difficulty identifying
Borderline-Group Method (cont.)

• 5 steps
  1. Select judges
  2. Define adequate, inadequate, and “borderline” levels of skills and knowledge tested
  3. Identify “borderline” test-takers.
  4. Obtain the test scores of the “borderline” test-takers
  5. Set cutoff score at the median test score of the borderline group. This is the score that divides the group exactly in half, i.e., half the members above and half below
Contrasting-Groups Method

• Test-takers divided into two contrasting groups
  • Qualified and unqualified
  • Based purely on judgment of their demonstrated knowledge and skills

• Judges pick a score and ask question – “Are the majority of these test-takers qualified or unqualified?”
  • Typically, those with higher scores are “qualified”

• Judges move down scale until “unqualified” outnumbers “qualified”
  • Judges decide where the cut-off will be, equal, higher, or lower
Contrasting-Groups Method (cont.)

• Seven steps:
  1. Select the judges.
  2. Define adequate and inadequate levels of knowledge and skills tested.
  3. Select the sample of test-takers whose skills will be judged. (Omit this step if you can get judgements of all test-takers.)
  4. Obtain the test scores and the judgements of the test-takers you have selected. Do not let the judges know the test scores.
  5. Divide the test-takers at each score level into “qualified” and “unqualified” groups on the basis of the judgements. Compare the percentage of test takers at each score level who are in the “qualified” group. (If you do not have several test-takers at each score level, combine score levels into larger intervals before you do this calculation.)
  6. Use a “smoothing” method to adjust the percentages you have computed.
  7. Choose a passing score on the basis of the “smoothed” percentage.
Contrasting-Groups Method (cont.)

- Data smoothing
  - Statistical method that uses a data frequency scatter plot and instead of connecting the dots in a jagged line, determines where the “smooth” curve should be based upon a curve that comes as close as possible the data points
  - Cut score is set at 50% qualified intersection
Up-and-Down Method

• Works best with small samples
  • Select one test-taker at a time
• Scores should have already been recorded
• 3 basic steps
  1. Select a test-taker with a test score near where you think passing should be. Make a personal assessment of this test-taker’s overall skills.
  2. Assuming the first test-taker was judged to be “qualified”, choose the next test-taker with a somewhat lower score. If the first test-taker selected was not judged to be “qualified” based on your observation, choose the next higher score and make an assessment of their skills.
  3. Repeat step 2 to choose the third test-taker based on the judgement you made of the second test taker. Continue in this fashion until you run out of test-takers.
# Hypothetical Data Using the Up-and-Down Method

## Test-Taker

<table>
<thead>
<tr>
<th>Test Score Out of Possible 20</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Numerical Score</strong></td>
<td>19</td>
<td>17</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
</tbody>
</table>

To find the passing score, average the scores of test-takers 4 through 16

**Passing Score = 13.6**
• The reason test-takers one through four were eliminated is because they were considered “qualified”.

• Test-taker five was the first indication of an “unqualified” test-taker based on the observation of the judge(s).

• When setting the final cut score, it is important to observe the points of several “reversals”, that is where the test-taker’s score goes the opposite direction of the previous one.

![Hypothetical Data Using the Up-and-Down Method](image)

```
<table>
<thead>
<tr>
<th>Test Score Out of Possible 20</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-Taker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>U</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Q</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Q</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td>U</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numerical Score</td>
<td>19</td>
<td>17</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
</tbody>
</table>

To find the passing score, average the scores of test-takers 4 through 16

Passing Score = 13.6
```
Body of Work Method

• Usually applied to K-12 testing
• Used to evaluate products that test takers produce
  • E.G. Essays
• Judges review all test questions and student responses
  • Responses placed in “response booklet”
  • Judges examine all and match demonstrated knowledge and skill to a performance level
Body of Work Method

• Usually three rounds of judgements
  • Training exercise
  • Range-finding stage
  • Pinpointing stage
Body of Work Method

• Training Stage
  • Judges examine 5-8 response booklets & match to performance levels
  • All share their results and discuss
  • Judges not told scores until the discussion period
Body of Work Method

• Range Finding Stage
  • After training stage judges review 30 or so response booklets
  • Categorize booklets to performance level without knowledge of scores
  • Judges discuss ratings and re-categorize if desired
  • Results can be used to determine roughly where each cut score will be
    • Based on distribution of scores of response booklets placed at each performance level
    • Example:
      • Booklets with scores <40% placed in basic level
      • Booklets with scores 41% - 60% placed in basic level by some and proficient level by others
      • Booklets with scores 61% - 75% placed in proficient level
      • Booklets with scores 76% - 85% placed in proficient level by some and advanced level by others
      • Booklets >85% placed in advanced level by all
Body of Work Method

• Distribution of example scores indicates cut score:
  • Between basic and proficient should be somewhere between 41% - 60%
  • Between proficient and advanced, 76% - 85%
Body of Work Method

• Pinpointing Stage
  • More booklets selected in ranges of scores that range finding round placed the cut scores
    • Typically 20 – 30 for each cut score
  • Booklets placed in one of the two relevant performance levels
  • Final recommendations of cut score made following group discussion
  • Actual booklets scores hidden from judges until end

• Final Cut Score
  • Find a point that best distinguishes between two adjacent performance levels
    • Similar to contrasting groups method

Copyright © 2017-2018 by Compass Consultants, LLC
“It is impossible to prove that a cut score is correct.”

ETS – A primer on Setting Cut Scores on Tests of Educational Achievement
Remember – calculated cut score may be modified by HR requirements and set higher or lower to meet organizational 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