Setting Performance Standards & Cut Scores
James B. Olsen, Senior Psychometrician & Accreditation Assessor
Interchangeable Terms

Cut scores = Passing Scores =
Passing Standards = Performance Standards
Session Objectives

1. Define terms.
2. Summarize professional standards for standard setting and cut scores.
3. Provide general framework and steps in standard setting.
4. Present and demonstrate alternate standard setting methods and procedures.
5. Summarize advantages and disadvantages of standard setting methods.
6. Give key professional references.
Defining Terms

**Standard Setting**: The process of deciding how much knowledge, skill and ability, usually expressed as a minimum, needs to be present to meet a determined standard.

**Cut Scores**: Test scores that divide a continuous distribution into categories and assign people with adjacent scores to different performance categories.

**Rubric**: A set of rules for scoring the responses on a performance or constructed-response test. Sometimes called a “scoring guide.”

**Performance Level Labels and Descriptors**: A label and descriptive statements of what examinees know and are able to do when classified as “certified” or “not certified.” or as “basic”, “proficient” or “advanced”.
Defining Terms

**Item/Task:** A test question, including the question itself, any stimulus material provided with the question, and the answer choices (for a multiple choice item) or the scoring rules or rubric (for a constructed-response item).

**Dichotomous Scores:** Item scores that are scored “1” for correct and “0” for incorrect. The only acceptable item or task scores are 0 and 1.

**Polytomous Scores:** Item scores that have multiple points such as 0, 1, 2, 3, 4 on a 5 point item or task scale. Acceptable scores could be 3, 4 or 3.6.

**Standard Error of Measurement:** A measure of how much test taker’s scores vary because of random factors, such as the date and time of the test, the particular selection of items on the test or form or the particular scorers who happened to score a test taker’s responses. The standard error of measurement is expressed in the same scale units as the scores themselves.
Standard Setting Decisions

• Pass/Fail
• Qualified/Unqualified
• Award a diploma/Do not graduate
• Grant Professional license or certification/Do not grant license or certification
• Place examinees into categories such as Basic, Proficient, Advanced
Minimally Qualified Candidate (MQC)

Cut Score

Not Yet Qualified | Qualified

Still Not Qualified | Minimally Qualified

Low Score | High Score
Standard Setting Impacts Validity

• Validity refers to a convergence of evidence (theoretical, logical and empirical) supporting the intended interpretation and use of test scores. It gives an evidence-based argument that supports the decisions and conclusions from test scores.

• The outcomes of standard setting (cut scores) directly inform score use and decision making.
Standard Setting Principles

• All cut score and standard setting methods involve judgment. There is no single correct cut score. Cut scores are constructed, not found.

• The cut score decisions can be informed by data but ultimately cut scores involve judgment-based decisions.

• Some people who should pass will fail (false negatives); some people who should fail will pass (false positives).

• Raising or lowering the cut score to reduce one type of error will increase the other type of error. Which classification error to minimize is a matter of stakeholder values.

• Decision makers need to take both types of error into account when setting cut scores.
Key Professional Testing Standards

Test scores and score scales should be developed in a way that supports the interpretations of scores for the proposed uses of tests. Test publishers and users should document evidence of fairness, reliability, and validity of test scores and score scales for their proposed use for all individuals in the intended examinee population. (5.0, in press)

When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented. (5.21, in press)

When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances. The judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way. (5.22, in press)
Key Professional Testing Standards

When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria. (5.23, in press)

When a validation rests in part on the opinions or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. (1.7) qualifications and experience; descriptions of procedures, training, instructions; group or independent judgments; level of agreement

Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (2.14)
Key Professional Testing Standards

When proposed interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented. (4.19)

When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria. (4.20)

When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules… (6.5)
Key Professional Testing Standards

When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performance or performance or test levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way. (4.21)

The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test. (14.17)
Common Elements in Standard Setting

1. Choose Standard Setting Method
2. Establish Performance Level Labels
3. Prepare Extended Performance Level Descriptions
4. Form Conceptualizations of the Minimum Qualified Candidate (MQC)
5. Select and Train Standard Setting Participants
6. Provide Feedback to Participants (Normative, Data and Impact)
Standard Setting Steps

1. Select a Standard Setting Method
2. Choose a Standard Setting Panel and Design
3. Prepare Descriptions of Performance Categories
4. Train Panelists to Use the Method
5. Collect Item and Task Ratings
6. Provide Feedback and Facilitate Discussion
7. Compile Panelist Ratings and Obtain Performance Standards
8. Conduct Panelist Evaluation
9. Compile Validity Evidence and Prepare Technical Documentation
Operational Steps

1. Select judges.
2. Teach judges about cut scores.
3. Define “borderline examinee.”
4. Train judges in use of standard setting method.
5. Run cut score study.
6. Set operational cut scores.
Operational Components

1. Help panelists gain a common understanding of the purposes and processes of setting standards.
2. Develop common agreement on the meaning of the policy definitions and desired performance levels.
3. Provide training on the selected standard setting process.
4. Implement the judgment ratings process and provide feedback on panelist judgments and candidate performance.
5. Evaluate the standard setting process.
6. Make final adjustments to panelist ratings based on the feedback and compute final cut scores and standard error confidence intervals.
7. Select exemplary items and skills to effectively communicate the cut score and performance standard.
Standard Setting Methods

I. Methods that involve review of test items and scoring rubrics. (Item Centered)
   – Angoff (Modified Angoff)
   – Extended Angoff
   – Bookmark and Item Mapping
   – Direct Consensus

II. Methods that involve review of candidates. (Examinee Centered)
   – Borderline Groups (Borderline Survey)
   – Contrasting Groups
Angoff and Modified Method

• Panel of experts (recommends cut scores)
  – Takes/Reviews Test
  – Defines what “qualified” means
  – Judges probability that a “qualified” test taker would know the correct answer to each multiple choice item.

• Decision makers set the final cut score

<table>
<thead>
<tr>
<th>No Chance</th>
<th>Even Chance</th>
<th>Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>0  .05  .10  .20  .30  .40  .50  .60  .70  .80  .90  .95  1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Unlikely</td>
<td></td>
<td>Very Likely</td>
</tr>
</tbody>
</table>
## Sample Angoff Data

### Modified Angoff Ratings: Round 2

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Rater ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>80</td>
<td>90</td>
<td>90</td>
<td>100</td>
<td>90</td>
<td>90</td>
<td>90.0</td>
<td>6.32</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>70</td>
<td>80</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>75.0</td>
<td>10.49</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>90</td>
<td>80</td>
<td>90</td>
<td>70</td>
<td>80</td>
<td>60</td>
<td>78.3</td>
<td>11.69</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>70</td>
<td>70</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>80</td>
<td>71.7</td>
<td>7.53</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>80</td>
<td>70</td>
<td>90</td>
<td>60</td>
<td>80</td>
<td>60</td>
<td>73.3</td>
<td>12.11</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>70</td>
<td>60</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>68.3</td>
<td>4.08</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>80</td>
<td>60</td>
<td>80</td>
<td>70</td>
<td>60</td>
<td>70</td>
<td>70.0</td>
<td>8.94</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>70</td>
<td>50</td>
<td>80</td>
<td>70</td>
<td>50</td>
<td>90</td>
<td>68.3</td>
<td>16.02</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>90</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>60</td>
<td>80</td>
<td>73.3</td>
<td>10.33</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>80</td>
<td>70</td>
<td>90</td>
<td>60</td>
<td>100</td>
<td>80</td>
<td>80.0</td>
<td>14.14</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>78.0</td>
<td>70.0</td>
<td>78.0</td>
<td>71.0</td>
<td>75.0</td>
<td>77.0</td>
<td>74.8</td>
<td>6.59</td>
</tr>
</tbody>
</table>

Note: Columns 7 to 13 deleted for Presentation Display
Extended Angoff Method

- Panel of Experts (recommends cut scores)
  - Takes/Reviews Test
  - Defines what “qualified” means
  - Judges the mean score that a “qualified” test taker would earn on each constructed response item/task.
- Decision makers set the final cut scores.
## Sample Extended Angoff Data

<table>
<thead>
<tr>
<th>Item-Round</th>
<th>Rater ID Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Means</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td>2.17</td>
<td>0.753</td>
</tr>
<tr>
<td>1-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.83</td>
<td>0.408</td>
</tr>
<tr>
<td>2-1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1.50</td>
<td>0.548</td>
</tr>
<tr>
<td>2-2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.17</td>
<td>0.408</td>
</tr>
<tr>
<td>3-1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td>2.00</td>
<td>0.000</td>
</tr>
<tr>
<td>3-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>2.83</td>
<td>0.408</td>
</tr>
<tr>
<td>4-1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.50</td>
<td>0.548</td>
</tr>
<tr>
<td>4-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>3.00</td>
<td>0.000</td>
</tr>
<tr>
<td>5-1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1.33</td>
<td>0.516</td>
</tr>
<tr>
<td>5-2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1.83</td>
<td>0.408</td>
</tr>
<tr>
<td>6-1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.50</td>
<td>0.548</td>
</tr>
<tr>
<td>6-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.83</td>
<td>0.408</td>
</tr>
<tr>
<td>7-1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.33</td>
<td>0.516</td>
</tr>
<tr>
<td>7-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>3.00</td>
<td>0.000</td>
</tr>
<tr>
<td>8-1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2.50</td>
<td>0.548</td>
</tr>
<tr>
<td>8-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>3.00</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Means R1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.00</td>
<td>2.38</td>
</tr>
<tr>
<td><strong>Means R2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.75</td>
<td>2.75</td>
</tr>
</tbody>
</table>
Bookmark Method

• Panel of Experts (recommends cut scores)
  – Takes/Reviews test
  – Defines what “qualified” means
  – Reviews test items presented in a booklet in order of their difficulty (Ordered Item Booklets, OIB)
  – Judges note where in the booklet a “qualified” test taker would likely know the test content up to that point, but not likely know the content beyond that point.

• Decision makers set the final cut scores.
Contrasting Groups Method

- Panel of Experts (recommends cut scores)
  - Takes/Reviews test
  - Defines what “qualified” means
  - Classifies (sorts) people based on their known levels of knowledge/skill into “qualified and “unqualified” categories

- People who were classified take the test
- Test scores of classified people are used to identify cut scores that best separates the two or more categories.
- Decision makers set the final cut scores.
Sample Bookmark Booklet

Bookmark Practice Test
U. S. History

Using the Performance Level Descriptor for Proficient, identify the most difficult item that a student who is just barely proficient would have a 2/3 chance of answering correctly. Enter its page number in the box below.

Bookmark
Page Number
Direct Consensus

- Panel of Experts (recommends cut scores)
  - Takes/Reviews test
  - Defines what “qualified” means
  - For each content cluster, panelists estimate the number of items the “qualified” candidate would answer correctly.
  - Panelists jointly discuss the results for content clusters and for the test as a whole.

- Decision makers set the final cut scores.
## Sample Direct Consensus Data

<table>
<thead>
<tr>
<th>Subtest Area (Number of Items)</th>
<th>Participants' Recommended Number of Items within a Subtest Area that the Just-Qualified Examinee should Answer Correctly</th>
<th>Subtest Area Means (SD)</th>
<th>Percent of Number of Items in Subarea</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (9)</td>
<td>7 7 6 7 7 7 6</td>
<td>6.75 0.46</td>
<td>75.0</td>
</tr>
<tr>
<td>B (8)</td>
<td>6 6 7 6 5 7 7</td>
<td>6.38 0.74</td>
<td>79.7</td>
</tr>
<tr>
<td>C (8)</td>
<td>5 6 7 6 6 6 5 6</td>
<td>5.88 0.64</td>
<td>73.4</td>
</tr>
<tr>
<td>D (10)</td>
<td>7 8 8 7 7 8 7 7</td>
<td>7.38 0.52</td>
<td>73.8</td>
</tr>
<tr>
<td>E (11)</td>
<td>6 5 7 7 7 8 8 6</td>
<td>6.63 0.92</td>
<td>60.2</td>
</tr>
<tr>
<td><strong>Rater Sums</strong></td>
<td><strong>31 32 35 33 32 35 34 32</strong></td>
<td><strong>33.00 1.51</strong></td>
<td><strong>71.7</strong></td>
</tr>
</tbody>
</table>

**Grand Mean (SD)**
Borderline Group (Borderline Survey)

- Panel of Experts creates Borderline Survey
  - Takes/Reviews test
  - Defines what “qualified” means
  - Creates survey items to identify and distinguish levels of competency related to each of the major test domain objectives. Recommend 10-15 survey items.
- Beta group of candidates takes the exam and borderline survey.
- Borderline Group identified through statistical analysis of survey results.
- Median or Mean score of Borderline Group is identified.
- Decision makers set the final cut scores.
<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angoff</td>
<td>Well researched; Ease of Implementation; not data intensive</td>
<td>Data entry; Sometimes difficult task for Panelists; Not efficient for setting multiple cut scores.</td>
</tr>
<tr>
<td>Extended Angoff</td>
<td>Generally applicable to all polytmous (multi-point) items and tasks.</td>
<td>Panelist task is cognitively challenging; Requires multiple rounds for convergence; Small test proportion is typically scored polytomously.</td>
</tr>
<tr>
<td>Bookmark</td>
<td>Efficient for multiple cut score; Same method applies to selected choice and constructed response items.</td>
<td>Needs pilot testing for IRT or ordering analysis; Requires good data and item span (small item difficulty gaps) for OIBS; Cut score analysis not transparent for panelists; Standards vary by RP values; Panelist task is more challenging for polytomous scored items (crossing multiple pages).</td>
</tr>
<tr>
<td>Method</td>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Borderline Groups (Beta &amp; Borderline Survey)</td>
<td>Candidate groups are based on independent evaluation criteria by survey or teacher/trainer judgments.</td>
<td>Lack of published literature on borderline survey. Survey results may correlate poorly or negatively with exam scores.</td>
</tr>
<tr>
<td>Contrasting Groups</td>
<td>Same method applies to selected response and constructed response items; based on real examinees or their products; Straightforward judgment</td>
<td>Make sure classifications are reasonable; Recruit test takers for contrasting groups; Requires stability of knowledge and technology.</td>
</tr>
</tbody>
</table>
Conclusions

There is no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their ultimate defensibility.

There is no sharp difference between those just below the cut score and those just above it.

Regardless of the cut score chosen, some examinees with inadequate skills are likely to pass and some with adequate skills are likely to fail. The relative probabilities of false positive and false negative errors will vary depending on the cut score chosen. (Draft Joint Standards, Cut Scores)
Conclusions

- Standard setting impacts test validity. Standard setting is an integral part of the test development and delivery process. The outcomes of standard setting (cut scores) directly inform score use, score interpretation, and decision making.

- Setting cut scores requires the involvement of policymakers, test sponsors, measurement professionals, and others in a multi-stage, judgmental process.

- Cut scores should be based on a generally accepted methodology and reflect the judgments of qualified people.
**Key References**


Key References


Contact Information

jamesbolsen@hotmail.com
Phone (801) 224-2200