Section IV
Additional Issues

This section includes some additional issues that are related to testing.
Many schools provide faculty with item analysis output following each multiple-choice examination. This output is an excellent source of information about an item and is useful in evaluating the quality of the item, as well as in evaluating the accuracy of the answer key.

The following are sample results from four items; each illustrates a common situation. The students taking the test were divided into a Hi group and a Lo group, based on their performance on the total test. If you have a small number of examinees, include the top 50% of the students in the Hi group and the bottom 50% in the Lo group. If you have a large number of examinees, you might include the top 25% in the Hi group and the bottom 25% in the Lo group.

Typically, item analysis output indicates the percentage of students in each group who selected each option. Often, it also includes some measure of item difficulty (e.g., the “p-value” or the proportion of students who answered the item correctly) and some measure of discrimination (e.g., a biserial or a point biserial). We recommend that attention be focused on the pattern of responses rather than on the difficulty level or discrimination index.
For each sample item below, the percentage of students selecting each option is shown. The total row shows the percentage of the total group who selected each option. For example, in Item #1, 1% of the Hi group selected Option A; 1% selected B; 91% selected C; 4% selected D; 1% selected E; and 2% selected F. In the same item, 20% of the Lo group selected Option A; 6% selected B, etc. The asterisk on Option B indicates that B was the purported correct answer.

**Item #1**

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B*</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi</td>
<td>1</td>
<td>1</td>
<td>91</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Lo</td>
<td>20</td>
<td>6</td>
<td>51</td>
<td>14</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>2</td>
<td>76</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

- p-value: 2
- discrimination index: -0.21

*Interpretation:* This is the typical pattern for an item that is miskeyed: if the answer is Option B, the item is very difficult and the discrimination index is negative. With a key of B, only 2% of the students answered correctly. The correct answer is almost certainly C, but a content expert should review the item to make sure. If the correct answer is C, the p-value becomes 76 and the discrimination index becomes 0.46 — these are both excellent from a statistical perspective, and there is no reason to make any changes in item text.

**Item #2**

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B</th>
<th>C*</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi</td>
<td>0</td>
<td>1</td>
<td>90</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Lo</td>
<td>0</td>
<td>1</td>
<td>60</td>
<td>25</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>1</td>
<td>74</td>
<td>12</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

- p-value: 74
- discrimination index: 0.33

*Interpretation:* 90% of the Hi group and 60% of the Lo group selected the correct answer. These are excellent overall statistics. You could rewrite A and B before you reuse the item because few students selected those options.
Interpretation of Item Analysis Results

Chapter 8

Item #3

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B</th>
<th>C*</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi</td>
<td>44</td>
<td>1</td>
<td>50</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Lo</td>
<td>20</td>
<td>15</td>
<td>21</td>
<td>22</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>7</td>
<td>34</td>
<td>14</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>

p-value: 34  
discrimination index: 0.30

Interpretation: 50% of the Hi group and 21% of the bottom group selected the correct answer. This is a very difficult item that is probably NOT OK. Too many of the Hi group selected Option A; the item may be poorly worded. Check Option A for “fairness.” Make sure Option A is not equally correct.

Item #4

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B</th>
<th>C*</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi</td>
<td>18</td>
<td>10</td>
<td>51</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Lo</td>
<td>24</td>
<td>24</td>
<td>21</td>
<td>25</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>17</td>
<td>34</td>
<td>22</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

p-value: 34  
discrimination index: 0.30

Interpretation: The Hi/Lo group breakdown on option C is identical to Item #3, but this item may be OK. In contrast to Item #3, those who don’t know the correct answer are widely spread across the various distractors. Of course, it would still be desirable to review options A, B and D for correctness and clarity.
Chapter 9  
Establishing a Pass/Fail Standard

Definitions and Basic Principles

Standards may be classified as either relative or absolute. A relative standard is based on the performance of the group taking the test. Examinees pass or fail depending upon how well they perform relative to other examinees taking the test. The following are examples of relative standards:

- Those scoring below 1.2 standard deviations below the mean will fail.
- The bottom 20 percent of the group will fail.

In contrast, an absolute standard does not compare the performance of one examinee with the others who are taking the test. Examinees pass or fail based only upon how well they perform, regardless of the performance of other examinees. All examinees could pass or all could fail. The following is an example of an absolute standard:

- Those answering less than 60 percent of the questions correctly will fail.

Unless there are strong reasons to fail a given number of examinees, an absolute standard (based on examinee performance) is preferred over a relative standard (based on a particular failure rate).

Basic Principles of Setting Standards

- Regardless of the procedure used, setting standards requires judgement. Setting standards will always be arbitrary, but need not be capricious.

- Unless there is a specific reason to fail a given number of examinees (eg, there are only a fixed number of slots available), a standard based on examinee mastery of exam content is preferred over a standard based on a particular failure rate.
• It is wise to involve multiple informed judges in the standard-setting process. Differences of opinion will occur, and use of multiple judges will reduce hawk/dove effects.

• Judges should be provided with data on examinee performance at some point in setting standards. Setting standards without such data may lead to uninformed standards and unreasonable results.

A helpful “how-to” reference on standard setting is:

Two Standard-Setting Methods Based on Judgements about Items

The Modified Ebel Procedure

• A group discusses the characteristics of the “borderline examinee”: an examinee whose skills are just good enough to allow him/her to pass.

• Judges categorize items as “Essential,” “Important,” or “Indicated.”

• Judges indicate the number of items in each category that a borderline examinee would obtain.

• The pass/fail standard is calculated as the percentage of possible points that a borderline examinee would obtain.
The Modified Angoff Procedure

- A group discusses the characteristics of a “borderline” examinee.

- For each item on the test, the judges estimate the percentage of borderline examinees who would answer the item correctly.

- The pass/fail standard for the test is the average of the percentages for the items.

Common Variations on the Angoff Procedure

- Judges may or may not be provided with the correct answers to questions.

- Judges may or may not be provided with information concerning the percentage of examinees who answered each item correctly.

- After a period of training, judges may continue to work as a group or may work individually.
Relative/Absolute Compromise Standards: The Hofstee Method

More recently, several “compromise models” have been developed that utilize the advantages of both relative and absolute standard-setting procedures. One of these methods, the Hofstee method, is described below.

1. Judges are asked to review a copy of the exam.

2. Judges then indicate the following values, which define acceptable standards:
   - Lowest acceptable percentage of failing examinees (minimum failure rate)
   - Highest acceptable percentage of failing examinees (maximum failure rate)
   - Lowest score which would allow someone to pass (minimum passing point)
   - Highest score required for someone to pass (maximum passing point)

3. After test administration, a curve showing the fail rate as a function of passing score is plotted. (In the figure shown, the curve extends from bottom left to top right.)

4. The four values obtained in #2 are drawn, forming a rectangle. Often the median values of the group of judges are used. In the example, the appropriate failure rate was judged to be between 0 and 20% (see horizontal lines); the appropriate pass/fail point was judged to be between 50 and 60% correct (see vertical lines).

5. A line is drawn on the diagonal from upper left to lower right. The point where this intersects the curve is the standard (ie, just above 55% correct in the figure).

A useful reference on compromise methods is:
Comments on a hodge-podge of topics related to testing are provided below. In general, the points made are speculative and based on anecdotal experience rather than evidence. That is, they reflect our biases rather than the results of research.

Multiple Station Exams (a.k.a. Practical Exams, Steeplechases, OSCEs)
Though logistically complex to set up and administer, these are very useful in the basic sciences, particularly to assess hands-on skills that cannot be measured with paper-and-pencil tests (eg, ability to use a microscope, to perform a laboratory procedure). In addition, reproduction of some kinds of material (eg, results of imaging studies, color pictorial material) is very expensive; in such situations, the multiple-stations approach can be used to reduce test administration costs.

Take-Home Exams
Take-home exams can be a substantial learning experience for students by stimulating them to read broadly and deeply on important topics. Unfortunately, students tend to produce tomes as answers, and it can be unclear if submitted answers represent the student’s own work. The same advantages can be gained by distributing (a superset of) test questions in advance, and administering (a subset of) questions as a timed test.

Open-Book Tests
Open-book tests can be a very good idea because of the impact on the kinds of questions that faculty prepare. For open-book tests, it is pointless to ask questions about isolated facts that can be looked up quickly on a single page of a text book, so test material developed for these tests tends to focus more on understanding of key concepts and principles in problem situations.
Frequent Short Quizzes versus Infrequent Tests
Infrequent testing makes each exam a major event; students may even stop attending class to prepare, and this seems undesirable. In addition, with infrequent tests, students may be unable to determine if they are studying the right material or learning in enough depth. Though it may be more time consuming for faculty, frequent testing reduces the importance of each individual exam and helps students to better gauge their progress. On the whole, frequent testing seems preferable, though students are likely to complain regardless of the approach adopted.

Keeping Tests “Secure” versus Permitting Students to Retain Them
Because tests can have a substantial “steering” effect on student learning, permitting students to retain test material can aid in focusing student attention on key topics, reinforcing curricular goals and course objectives (assuming test materials reflect these). However, preparation of good exam questions is very time consuming, and, over time, the quality of test material can deteriorate if faculty have to develop new test materials each time a course is taught. The best approach may be to make sample good-quality test material available in order to influence student learning, but maintain a bank of “secure” questions for repeated use, keeping in mind that security is likely to be poor, since students commonly memorize questions and reproduce them for each other.

Use of Cumulative Tests
Cumulative tests that hold students responsible for all material presented to date encourage attention to inter-relationships among topics, particularly if test questions require understanding of both old and newly presented topics. Use of tests that cover only material presented since the previous test encourages students to study topics in isolation; relationships among topics from different units may be missed. Since students can do badly on a series of tests because they never master basic material, this approach can also motivate students to remediate weaknesses.

Use of Integrative, Cross-Course Tests
Like the use of cumulative tests, integrative cross-course tests encourage students to see inter-relationships among disciplines and topics; this should be very helpful for long-term retention and for application of basic science knowledge to clinical situations. Generally, faculty from both basic science and clinical departments are needed for preparation of such exams. While time consuming, this joint effort may result in better test material as well as useful discussion among faculty of what material should be included in the curriculum.