SCALED SCORES
Their Rationale, Derivation, and Interpretation

Many state regulatory boards for many professions must cite a specific passing standard in their practice statute. This standard must be the same for different tests, and for different administrations of the same test. Yet, the psychometric characteristics of the tests, particularly their difficulty levels, may vary widely. Since difficulty levels partially determine pass-fail standards, and as noted, difficulty levels vary, a statute that mandates a specific pass-fail standard creates a dilemma.

Scaling is the statistical procedure that resolves this dilemma. Unfortunately, this resolution creates misunderstanding about what scaling is, and what the scores represent. This article discusses the philosophy and methodology of scaling, and provides an example to facilitate interpretation of scaled scores. The purpose is to promote a better understanding of what scaling is, and what it is not.

Philosophy
There is great uncertainty about why practice statutes reference specific pass-fail standards. However, there is a large number and wide variety of licensed professions whose practice statutes indeed cite a specific passing standard, typically “75.” Perhaps, by referencing a constant standard, regulators felt that standards would not be vulnerable to manipulation or to irrational changes. Nonetheless, standards do vary, and in a rational, logical manner, and if 75 were interpreted as a percentage score, the solid majority of students would fail the typical exam. This failure would not be the result of poor candidate preparation but rather, because of an unrealistically high standard.

Scaling provides consistency in reporting passing standards, while allowing for the standards to vary. Scaling also enables passing standards to be established at more realistic levels. Because it is able to meet these conflicting demands, scaling (of some sort) is used by many, if not most, national boards of fully licensed professions for reporting scores. Here is how the NBEO computes scaled scores.

Methodology and Interpretation
Procedurally, scaling is simply a conversion. For each test, the conversion uses the actual pass-fail score in a pair of linear equations. The actual pass-fail score is the minimum number of test items that must be answered correctly to pass. The methodology used in determining the actual pass-fail score is described in another article on the NBEO web site.

For NBEO examinations, the part scores and section scores found on the front side of National Board score reports, are scaled scores. In scaling Part I (Basic Science), Part II (Clinical Science), or Part III (Patient Care), the pass-fail score is scaled to be reported as 300. In addition, the maximum possible score (i.e., 100% correct) is scaled to be reported as 900, and the minimum possible score (i.e., 0% correct) is scaled to be reported as 100. Passing scores lie between 300 and 900, inclusive. Failing scores lie between 100 and 299, inclusive.

In scaling section scores, the pass-fail score is reported as 75. However, please note that with the exception of the TMOD® section, NBEO pass-fail outcomes are based on the overall Part only. Passing the TMOD® section is an NBEO requirement because of expanded scope of practice considerations. All other section standards are established as a courtesy to those state boards whose practice statute requires that candidates pass individual sections within a part. In converting section scores, the pass-fail score for each section is scaled to be reported as 75. The maximum attained score (i.e., 100% correct) is scaled to be reported as 99, and the minimum possible score is scaled to be reported as 0 (zero). Passing scores lie between 75 and 99, inclusive. Failing scores lie between 0 and 74, inclusive.
Printed below are part and section data from a Part II (Clinical Science) test, accompanied in the last column by the corresponding scores of a real candidate. This candidate passed the part, and attained a scaled score of 75 or higher on every section except Perceptual Conditions. The equations used to scale her scores are shown below. However, in order to use the equations, several statistics must be known. These statistics follow the table.

<table>
<thead>
<tr>
<th>Examination</th>
<th>Highest Raw Score Possible* (Equal to the number of Items Scored)</th>
<th>Raw Pass/Fail Cutoff Score</th>
<th>Raw Score For This Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part II</td>
<td>423</td>
<td>269</td>
<td>281</td>
</tr>
<tr>
<td>Sections</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systemic Conditions</td>
<td>67</td>
<td>38</td>
<td>52</td>
</tr>
<tr>
<td>Ocular Disease/Trauma</td>
<td>178</td>
<td>115</td>
<td>116</td>
</tr>
<tr>
<td>Refractive ... Conditions</td>
<td>120</td>
<td>78</td>
<td>85</td>
</tr>
<tr>
<td>Perceptual Conditions</td>
<td>33</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>Public Health</td>
<td>13</td>
<td>88</td>
<td>98</td>
</tr>
<tr>
<td>Legal and Ethical Issues</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMOD®**</td>
<td>90</td>
<td>58</td>
<td>59</td>
</tr>
</tbody>
</table>

* Statistics are based on 423 scored items. Of the 435 items administered, 12 were considered flawed and therefore were deleted from scoring.

** TMOD® items are embedded within (and scored within) the Ocular Disease/Trauma section.

First, the Part scores are scaled, using the standard formula: \( y = ax + b \). To calculate the scaled score for the Part, candidates whose Part II score is at or above the pass-fail score (i.e., 269 or above) would use the following equations, based on a maximum scaled score of 900 being equal to the highest raw score possible, and a pass-fail scaled score of 300 being equal to the actual raw pass-fail score.

\[
900 = 423x + b \\
300 = 269x + b
\]

Solving these two equations yields the following values:

\[
x = 3.896 \\
b = -748.052
\]

With these variables calculated, all passing scores (i.e., raw scores greater than or equal to 269, represented by \( x \)) are scaled with the formula: \( 3.896x - 748.052 \). For this candidate, this equation yields a Part II scaled score of 347.

For candidates who failed Part II, the following scaling equations were used:

\[
299 = 268x + b \\
100 = 0x + b
\]

Solving these two equations yields the following values:

\[
x = 0.743 \\
b = 100.000
\]

With these variables calculated, all failing scores (i.e., raw scores less than 269, represented by \( x \)) are scaled with the formula: \( 0.743x +100 \). If this candidate had answered 268 items correctly, this equation would have produced a Part II scaled score of 299.

Revised 10/6/03
With the Part scaling complete, the sections are scaled. Section scaling uses the same 2-equation structure and solution, but with different values for the pass-fail cutoff score (i.e., 75, rather than 300). Candidates whose Systemic Conditions score is at or above the pass-fail score (i.e., 38 or above, for state board purposes) would use the following equations, based on a maximum scaled score of 99 being equal to the highest raw score possible, and a pass-fail scaled score 75 being equal to the actual raw pass-fail score.

\[ 99 = 67x + b \]
\[ 75 = 38x + b \]

Solving these two equations yields the following values:

\[ x = 0.828 \]
\[ b = 43.524 \]

With these variables calculated, all passing scores (i.e., raw scores greater than or equal to 38, represented by \( x \)) are scaled with the formula: \( 0.828x + 43.524 \). For this candidate, this equation yields a Systemic Conditions scaled score of 87.

For candidates who failed Systemic Conditions, the following scaling equations were used:

\[ 74 = 37x + b \]
\[ 0 = 0x + b \]

Solving the two above equations yields the following values:

\[ x = 2.000 \]
\[ b = 0.000 \]

With these variables calculated, all failing scores (i.e., raw scores less than 38, represented by \( x \)) are scaled with the formula: \( 2.000x + 0.000 \). If this candidate had answered 37 items correctly, this equation would have produced a Systemic Conditions scaled score of 74. Each of the other sections would follow the same scaling procedure, but with different values, reflecting a different number of scored items and pass-fail standard.

In interpreting scaled scores, two common misconceptions should be avoided. First, recognize that the Part pass-fail scaled score of 300 does not represent one-third or 33% of 900. In the example above, the scaled cutoff score of 300 was equal to 63.6% (269/423). Second, a section scaled cutoff of 75 is not likely to be 75%. For the four sections in Part I and the six regular sections in Part II, the cutoff score rarely reaches 75%. For Part III, the cutoff score for the clinical skills section has always exceeded 75%. Only the PAM (Patient Assessment and Management) test in Part III has a cutoff score that is likely to consistently equal 75%, but that cutoff changed to 74% in 2002 to correspond to a slight change in the test format.

**Beyond Scaled Scores**

Despite the value of scaled scores, there is no compelling reason for withholding raw (i.e., number right) and/or percentage scores. These latter scores are more meaningful to candidates with respect to assessing how much one really "knows." With this in mind, the National Board discloses these scores by publishing the section and part national statistics with raw and percentage score equivalents of the scaled statistics. All candidates also receive a personal supplemental performance analysis on side two of their score report noting their raw scores (i.e., number correct) for each section, and each major content area within each section.

Revised 10/6/03
The number correct in each content area is accompanied by the number of items scored, enabling candidates to compute their percentage scores and to compare their areas of relative strength and weakness between sections, and within a section. For example, consider another candidate who failed the test. This candidate attained a percentage score of 81% in Refractive Conditions and 59% in Perceptual Conditions. Though seemingly stronger (and above a scaled score of 75) in Refractive Conditions, he may have scored 9 of 11 (82%) in area 4 (low vision), but only 2 of 13 (15%) in area 6 (eye movement anomalies). Therefore, even in an area of relative strength, there is a relatively weak area identified for further study. Similarly, this candidate’s weak score in Perceptual Conditions may be concentrated in one content area. For example, he may have scored only 3 of 14 (21%) in area 1 (child development anomalies) but 16 of 19 (84%) across the other three content areas. This personal supplemental performance analysis can be particularly useful in helping to guide the preparation of candidates who need to retake the test.

National Board examinations in any profession are designed to assess, not teach. However, the outcome data can be of substantial value, and there is nothing to be gained by suppressing or obscuring this information, as long as it is not misused. Hopefully, this explanation of scaling will help candidates better understand their score reports, and optimize the educational usefulness of the information.