# National Board of Medical Examiners ®
## Subject Examination Program

### Analyzing Performance on Subject Tests
### Medicine

Many faculty have asked for assistance with determining appropriate scores for passing Subject Tests and for obtaining honors. A general approach is presented here that may be used by course and clerkship directors to perform such analyses.

A standard is a value that answers the question "How much is enough?" Standards are set in nearly every industry and profession to protect the general public. Standard setting may be as disparate as determining the minimum amount of protein required in a school lunch, to determining the maximum speed limit on a highway or the minimum passing score on an examination. It is important to recognize that, because judgment is always involved in the standard setting process, in a sense all standards are somewhat arbitrary. On the other hand, it is also important to note that the standards should not be capricious.

**Background: Relative and Absolute Approaches**

Standards may be classified as either *relative* or *absolute*. A *relative standard* is based on the performance of the group taking the same exam. Examinees are classified (e.g., Pass/Fail, Honors) depending upon how well they perform relative to other examinees taking the exam. The following are examples of *relative* standards:

those scoring 1.2 standard deviations or more below the mean will fail
the top 10 percent of the group will achieve Honors

In contrast, an *absolute standard* does not compare the performance of one examinee with the others who are taking the exam. Examinees are classified based only upon how well they perform, regardless of the performance of other examinees. In theory, all examinees could meet the standard or all could not. The following are examples of *absolute* standards:

those answering less than 60 percent of the questions correctly will fail
those answering at least 85% of the questions correctly will achieve Honors

For several reasons, use of absolute standards has substantial intuitive appeal. First, it seems more equitable to base pass/fail decisions on the quality of an individual examinee's performance; it does not seem reasonable for classification decisions to be determined by the strengths and weaknesses of other examinees taking the same test. Second, if all examinees perform well, it seems reasonable that all should pass, rather than predetermining that a specific percentage of examinees will fail. Third, it is conceptually appealing to think of a standard as reflecting the minimum level of performance required to practice safely, continue with training, achieve Honors level, etc.

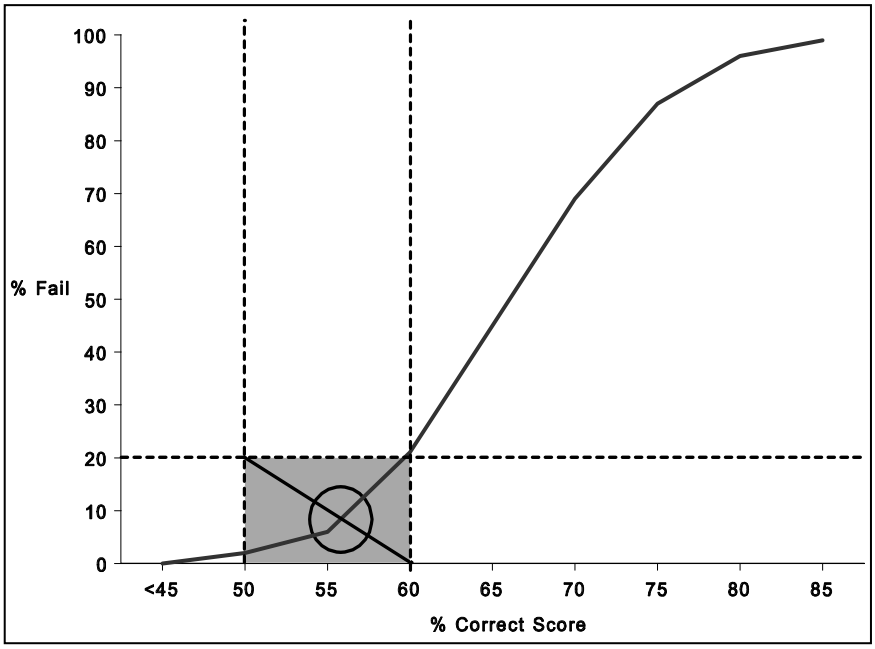
However, from a practical perspective, it is often difficult to agree on the absolute level of performance that should serve as a pass/fail point, particularly in advance of test administration. Often, test users have an intuitive sense of the overall quality of an examinee group and strong ideas about the rough proportion that should pass and fail. In effect, relative standards capitalize on test users' knowledge of the examinee group to calibrate the standard that is used. In many situations, use of relative standards is a very reasonable alternative.

# National Board of Medical Examiners ®
## Subject Examination Program

### A Relative/Absolute Compromise Approach: The Hofstee Method

A recent innovation in standard setting utilizes "compromise models", which 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 to require of 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 shown, 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 percent 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 percent correct in the figure).




A useful reference on compromise methods is: de Gruijter D. Compromise models for establishing examination standards. *Journal of Educational Measurement.* 1985;22:263-269.

A helpful "how-to" reference on standard setting is: Livingston SA, Zieky MJ. (1982) *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton: Educational Testing Service.

# National Board of Medical Examiners ®
## Subject Examination Program

### Compilation of Recent Analyses by Internists

Clerkship directors nationally in Medicine provided their own judgements regarding standards, and we have compiled and are now disseminating this information. Clerkship directors provided an opinion regarding the minimum and maximum percentage of students who should pass; as well as the lowest score that should allow someone to pass and the highest score that should be required of someone to pass. Each clerkship director was also asked to provide opinions regarding the maximum and minimum percentage of students who should achieve Honors level, as well as the minimum and maximum scores required to obtain Honors using the Hofstee method described on the preceding page.

A summary of the findings on the pass/fail and honors standards are provided below. Please note that results are computed on the **Subject Test Score** scale (the one reported in the first column of the **Roster of Scores**, with a mean of 70 and a SD of 8). This scale is used because it is an equated score; scores from one form of the test are comparable to scores on other forms of the test. This is close to (but not exactly the same as) a percent correct score.

The data shown below represent a compilation of the opinions of the 50 internists who participated in the study. These data are provided for your information. It is, of course, your decision whether or not you want to use these data in determining pass/fail or honor's standards in your clerkship.

**Medicine Results: Based on Responses from 50 internists**

**Minimum score for passing:**
- Mean: 60
- Range: 53 to 64
- Most values were between 58 and 62 (mean +/- 1 SD)

**Minimum score for honors:**
- Mean: 82
- Range: 77 to 96
- Most values were between 79 and 85 (mean +/- 1 SD)

10/09/97

# National Board of Medical Examiners ®
# Subject Examination Program

## Score Interpretation Guide
## NBME® Medicine Subject Test

NBME Subject Tests provide medical schools with a tool for measuring students' understanding of the clinical sciences. While Subject Tests are designed to be broadly appropriate for end-of-clerkship assessment, course objectives vary across schools, and the congruence between Subject Test content and clerkship objectives should be considered in the interpretation of test scores and in the determination of grading standards. NBME neither sets nor recommends a "passing" score. Generally, Subject Test Scores should be used in conjunction with other indicators of student performance in the determination of grades.

***Subject Test Scores***

The Roster of Scores reports a Subject Test Score for each examinee. These scores are scaled to have *a mean of 70 and a standard deviation of 8* for a group of approximately 10,000 first-time takers from 80+ schools who took the Medicine Subject Test as a final clerkship exam following rotations during the 1993-94 academic year. As a result, the vast majority of scores range from 45 to 95, and although the scores have the "look and feel" of percent-correct scores, they are not. This scale provides a useful tool for comparing the scores of your students with those of a large, nationally representative group taking the Subject Test as an end-of-clerkship assessment.

***Precision of Scores***

Measurement error is present on all tests, and the standard error (SE) provides an index of the (im)precision of scores. The SE is approximately 3 points for the Subject Test Scores. The SE indicates how far the score an exmainee earns on the exam is likely to vary from the examinee's "true" proficiency level. Like the standard error for a diagnostic laboratory study, the SE is expressed on the same scale as the test scores and can be used to construct confidence intervals around the scores. For example, if a student receiving a Subject Test Score of 60 were tested repeatedly with similar exams, 95% of the scores received should fall between 54 and 66 (60 plus/minus two times an SE of 3). While this level of imprecision may seem large, NBME Subject Tests provide scores that are, in general, more precise than tests developed by local faculty. Scores on course exams are not as precise as measurements in the biomedical sciences; small differences are not meaningful and should not be overinterpreted.

***Frequency Distribution of Scores***

If two or more students were tested, a Frequency Distribution is provided. The distribution shows the number (Count) and percent of students with each score, together with the cumulative frequency and percent. Summary information, including the mean, standard deviation, and the highest and lowest scores for the students tested, is also provided with the Frequency Distribution and the Roster of Scores.

## *Norms for Examinee Performance*

The table below provides norms to aid in the interpretation of student performance. These norms reflect the performance of 11,248 students from U.S. and Canadian medical schools who took the Medicine Subject Test as a final clerkship exam for the first time during the 2001-2002 academic year. The norms demonstrate the performance of examinees across the entire academic year and by quarterly testing periods. These norms allow you to compare your students' Subject Test scores with the performance of the group described above.

Quarterly norms have been provided because it is common knowledge that scores in some clerkship exams are progressively higher for students of equivalent ability who take the relevant rotation later in the academic year. For example, a percentile rank corresponding to a score of 75 for Quarter 1 is 65; in Quarter 4 the percentile rank for this score is 54. This information may have particular relevance to schools that have used the norm table in the development of grading guidelines.

For most schools the performance of examinees who took a Medicine Subject Test within the indicated months are represented in the following quarters:

Quarter 1: August, September & October

Quarter 2: November, December & January

Quarter 3: February, March & April

Quarter 4: May, June & July

The mean and standard deviation (SD) of this group for the Medicine Subject Test scores across the entire academic year and by quarter are as follows:

| | Mean | SD |
|---|---|---|
| **Total Year:** | 74.0 | 8.3 |
| **Quarter 1:** | 72.6 | 8.1 |
| **Quarter 2:** | 73.4 | 7.9 |
| **Quarter 3:** | 74.7 | 8.2 |
| **Quarter 4:** | 75.0 | 8.7 |

To use the table, locate a student's Subject Test score in the column labeled "Score" and note the entry in the adjacent column labeled "Percentile Ranks" for the Total Year or Quarterly testing period of interest.

For example, if a student's score is 63, the corresponding percentile rank entry of 10 for the Total Year indicates that 10% of the students taking the Medicine Subject Test during the 2001-2002 academic year are projected to have scores at or below 63.

Similarly, if a student's score is 63, the corresponding percentile rank entry of 8 for Quarter 4 indicates that only 8% of the of students taking the Medicine Subject Test during the months of May, June and July of the 2001-2002 academic year are projected to have scores at or below 63.

| | ***Percentile Ranks*** | | | | |
|---|---|---|---|---|---|
| **Score** | **Total Year** (n=11,248) | **Quarter 1** (n=2,796) | **Quarter 2** (n=2,524) | **Quarter 3** (n=3,133) | **Quarter 4** (n=2,795) |
| 93 or above | 99 | 99 | 99 | 98 | 97 |
| 92 | 98 | 99 | 99 | 98 | 97 |
| 91 | 98 | 99 | 99 | 98 | 96 |
| 90 | 97 | 98 | 98 | 97 | 96 |
| 89 | 97 | 97 | 97 | 96 | 95 |
| 88 | 95 | 96 | 96 | 95 | 94 |
| 87 | 94 | 96 | 96 | 93 | 92 |
| 86 | 93 | 95 | 95 | 92 | 90 |
| 85 | 91 | 93 | 94 | 90 | 88 |
| 84 | 89 | 92 | 91 | 88 | 86 |
| 83 | 87 | 91 | 89 | 85 | 83 |
| 82 | 84 | 88 | 87 | 82 | 80 |
| 81 | 81 | 86 | 84 | 79 | 77 |
| 80 | 78 | 83 | 81 | 76 | 73 |
| 79 | 75 | 80 | 78 | 72 | 70 |
| 78 | 72 | 78 | 75 | 68 | 66 |
| 77 | 68 | 75 | 71 | 64 | 63 |
| 76 | 64 | 71 | 68 | 61 | 59 |
| 75 | 59 | 65 | 62 | 55 | 54 |
| 74 | 53 | 60 | 56 | 49 | 48 |
| 73 | 49 | 56 | 52 | 45 | 45 |
| 72 | 45 | 52 | 48 | 41 | 41 |
| 71 | 41 | 48 | 43 | 37 | 36 |
| 70 | 34 | 40 | 36 | 30 | 31 |
| 69 | 30 | 36 | 32 | 27 | 27 |
| 68 | 26 | 32 | 27 | 23 | 24 |
| 67 | 22 | 26 | 23 | 19 | 20 |
| 66 | 19 | 23 | 20 | 16 | 17 |
| 65 | 15 | 19 | 15 | 12 | 13 |
| 64 | 12 | 16 | 13 | 10 | 10 |
| 63 | 10 | 12 | 10 | 8 | 8 |
| 62 | 8 | 10 | 8 | 7 | 6 |
| 61 | 6 | 7 | 6 | 5 | 5 |
| 60 | 5 | 6 | 5 | 4 | 4 |
| 59 | 3 | 5 | 3 | 3 | 3 |
| 58 | 3 | 3 | 2 | 2 | 2 |
| 57 | 2 | 2 | 2 | 2 | 2 |
| 56 | 1 | 2 | 1 | 1 | 1 |
| 55 | 1 | 1 | 1 | 1 | 1 |
| 54 and below | 1 | 1 | 1 | 1 | 1 |