Setting the Final Cut Scores

Overall Evaluation of the Candidate
The evaluation of an international candidate’s potential academic performance should be based on all available relevant information, not solely on the Test of English as a Foreign Language™ (TOEFL®) scores. General communicative effectiveness in English is only one of many qualities necessary for successful academic performance. Other qualities may include, for example, command of subject matter, interpersonal skills, and interest in the discipline. The TOEFL test does not, of course, provide information about aptitude, motivation, command of subject matter, or cultural adaptability, all of which may have significant bearing on a student’s academic success.

The new TOEFL Internet-based test (iBT) measures an individual’s English communication skills in reading, listening, speaking, and writing. Each examinee’s score report provides the four section scores on a 0-30 score scale and the total score on a 0-120 scale. ETS recommends that when making an admission decision based on TOEFL iBT scores, you consider the examinee’s profile of section scores rather than relying solely on the combined total score. There are many different combinations of section scores that would result in the same total score, but each combination represents a different profile of abilities across the four language modalities.

Considerations to Make When Setting the Final Cut Scores
It is consistent with acceptable standard-setting practice that the decision makers consider not only the panel’s recommendations for the cut score, but other important information such as different needs across departments, the number of applications for admission, and available information about how candidates, collectively, have performed on the test. These issues are detailed below.

Different departments/units have different needs: Not all departments require the same kinds of academic tasks, and not all academic tasks require the same level of performance across different aspects of English language ability. For example, in some institutions, first-year undergraduate students may primarily be attending lectures and taking multiple-choice examinations. It might be prudent to use the panel’s recommended cut scores for the Reading and Listening sections, or even increase them, if that seems appropriate. In this context, it might be acceptable to lower the cut scores for Writing and Speaking (especially if the institution provides support classes for ESL writing). An entirely different profile of skills might be deemed minimally acceptable for graduate school applicants: All graduate applicants might require a certain total score, but individual departments could require different profiles of section scores, depending on the specific language skill needs. One obvious example of this is requiring a higher Speaking cut score for first-year graduate students who wish to be Teaching Assistants (TAs). A special panel was convened to establish Speaking cut scores for International TAs. (See “Panel Recommendations for Scaled Cut Scores.”)

Number of applications for admission: If there is a shortage of entry-level candidates, it may be desirable to set a lower cut score. However, it is also important to consider the impact of such a decision. The goal of setting cut scores on the TOEFL test is to identify applicants who are non-native speakers of English who can communicate in English adequately.

Before setting a lower cut score requirement, one might consider questions such as:

- Can the university provide adequate remediation/English language support for those students who might need to improve their communication skills?
• Will the decision disadvantage students with inadequate communication skills if English language support is not available?

Before setting a higher cut score requirement, one might consider questions such as:

• Will the decision discourage international candidates from applying to the institution/department?
• Will the decision lead to rejection of otherwise qualified individuals who could make valuable contributions to the institution/department?

Performance on the test: The tables on pages 4-10 in “Score Comparison Percentiles” compare total and section scores on iBT (Internet-based) TOEFL to CBT (computer-based) TOEFL and PBT (paper-based) TOEFL. The comparisons between CBT and PBT are based on a 1998 study; comparisons between the new TOEFL test and CBT TOEFL are based on the 2004 field study (2,720 candidates from 30 countries). The information in the tables may be used as a guideline for how a candidate might perform on each of the different versions of the TOEFL test. When establishing your final cut scores, you may want to consider how those scores on TOEFL iBT line up against scores on the other versions of TOEFL. For example, if your institution currently accepts candidates with a total CBT score of 213, that would correspond to a 79 or 80 on TOEFL iBT. If the current CBT score has, in your institution’s judgment, been doing an acceptable job of identifying English-proficient candidates, you might want to consider setting your final cut score on TOEFL iBT at around 79. It is important to note, however, that these are score comparisons of the test versions, not exact score translations. There is no guarantee, for example, that a candidate who scored a 213 on the CBT would score a 79 or 80 if he or she took TOEFL iBT. We recommend that you use the tables as “rules of thumb” and not as precise indicators.

Tables showing percentile ranks for scores on the 2004 TOEFL iBT field study appear on pages 11-14 in “Score Comparison Percentiles.” Values are included for the total TOEFL iBT test and for the Reading, Listening, Writing, and Speaking sections. Basic statistics are also included, such as the mean and standard deviation. These tables show the percentage of students who earned a score equal to or below a certain value. For example, if you look at the Total Scale Score and Percentile Rank columns on page 11, 89% of the field test population (in the “Percentile Rank” column) earned a total TOEFL score of less than or equal to 100 (in the “Total Scale Score” column); 88.3% earned a total score of less than or equal to 99. These tables are useful in that they provide some insight into how your candidates may perform at different cut scores. The same caution advised above, however, needs to be exercised here: these percentiles are based on the 2,720 students who participated in the field test; your candidates may perform differently from this population. The tables on pages 4-5 show how the field study population performed on TOEFL CBT. This information sheds additional light on the proficiency of the field test population. This will help you interpret the percentile data and help inform your decision as to where to set your final cut scores.

It is also advisable to consider the standard error of measurement (SEM) when deciding on the final cut scores. The SEM is an index of the degree of uncertainty associated with a test score. A test score – any test score – is an approximation of what a candidate “truly” knows or is “truly” able to do. The less a test score is influenced by extraneous factors – factors not directly related to what was intended to be measured – the better a test score will approximate a candidate’s “true” score. Extraneous factors may relate to the candidate, the testing situation, the test content, and various combinations of these factors. One example of an extraneous factor related to the candidate is fatigue. A very tired candidate on the day of the test is not likely to do his or her best on the test; the test score for this candidate, therefore, may not be a very good approximation of what he or she “truly” knows or is “truly” able to do. There is, by definition, a 68% likelihood that a candidate’s “true” score (the score obtained without any extraneous
influences) is within one SEM of his or her test score; a 95% likelihood that a candidate’s “true” score is within two SEMs of his or her test score.

The SEMs based on the field study of TOEFL iBT appear below.

<table>
<thead>
<tr>
<th></th>
<th>1 SEM</th>
<th>2 SEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>2.29</td>
<td>4.58</td>
</tr>
<tr>
<td>Listening</td>
<td>2.31</td>
<td>4.62</td>
</tr>
<tr>
<td>Speaking</td>
<td>2.30</td>
<td>4.60</td>
</tr>
<tr>
<td>Writing</td>
<td>3.06</td>
<td>6.12</td>
</tr>
<tr>
<td>TOTAL</td>
<td>5.28</td>
<td>10.56</td>
</tr>
</tbody>
</table>

For example, if a candidate has a Reading score of 20, we can say that we are 68% confident that his or her “true” Reading ability is between 17.71 and 22.29 – 20 plus or minus one SEM (20 – 2.29; 20 + 2.29). We can say that we are 95% confident that his or her “true” Reading ability is between 15.42 and 24.58 – 20 plus or minus two SEMs (2 x 2.31 = 4.58).

The SEM for the total test is 5.28 scaled score points. If a candidate has a total score of 70, we are 68% confident that his or her “true” English language ability as reflected by total score is between 64.72 and 75.28.

The SEMs may be used to inform decisions about raising or lowering the recommended cut scores.

**Validation of the Cut Score Recommendations**

Once TOEFL iBT cut score recommendations for admissions have been set, institutions may consider conducting one or more studies to validate the use of the test and to evaluate the appropriateness of the cut scores.

One approach to validate the TOEFL iBT admission standards is to have admitted international students rated by their faculty members on their English language ability. If the majority of the students are rated as having adequate language ability for academic studies, it provides some supporting evidence for the appropriateness of the TOEFL iBT standards used.

**Periodic Reviews of the Cut Score Recommendations**

The distributions of TOEFL test takers’ scores may positively shift after the TOEFL iBT launches in September 2005, as the new TOEFL test is expected to influence English language learning and instruction. Institutions may wish to revisit their initial cut score recommendations on the TOEFL iBT test after it has been in place for a few years; it is also advisable to review the cut scores on a regular basis thereafter (for example, every three years) to evaluate how well the current set of cut scores are informing admissions decisions.

**Questions You May Want to Consider Before Setting the Final Cut Scores**

What cut scores does our university currently require for the computer-based (CBT) and paper-based (PBT) versions of the TOEFL test? How many applicants, on average, meet these cut scores?

Look up the comparable new TOEFL iBT scores for your current cut scores. How do the panel recommendations compare to these scores?
Is the language proficiency of the students we are currently accepting sufficient for them to cope with first year studies? If not, which of their language skills (Reading, Listening, Writing, Speaking) seem to be insufficient for coping with first-year studies?

Does our institution or department currently look at an applicant’s individual section scores or only the total test score? Do we currently look at an applicant’s score profile? (e.g., total CBT score of at least 213 but no section score below 20)

Are we willing to consider use of section/profile scores? If so, what kind of profile makes sense for our university? Can we have different profiles (section cut scores) for different departments?

How important is speaking ability to us? Would we want to not accept students who have adequate reading, listening, and writing skills because of inadequate speaking skills?