国際教育交換協議会（CIEE）日本代表部
作成資料

*TOEFL iBT®*テストスコアの CEFR マッピングに関して

本資料は、TOEFL テスト主催団体である Educational Testing Service（ETS）が発行している 2 つのリサーチレポートの一部を TOEFL テスト日本事務局である当協議会で取りまとめたものです。詳細については必ず原文（英語）にて確認ください。

---

1. TOEFL iBT テストスコアの CEFR レベルマッピング方法について（2006 年 10 月）

参照：”Linking English-Language Test Scores onto the Common European Framework of Reference: An Application of Standard-Setting Methodology”

2006 年 10 月 10 日～13 日の 4 日間で、23 名のパネリストが TOEFL iBT テストスコアと CEFR レベルとのマッピング作業を行った（①Reading と Listening は **Modified Angoff Method**、②Speaking と Writing は **Modified examinee paper selection method**）。

パネリスト構成

ETS ヨーロッパの言語スペシャリストが、16 か国、23 名にマッピング作業に参加するように依頼。23 名のパネリストは、英語教授法、英語学習、英語試験の専門家で、CEFR、TOEFL テスト等に精通している。彼らは、TOEFL テストを活用しているヨーロッパ各国を代表し選出された。

CEFR レベル内容を理解する （7～8 ページ）

カットスコア作業にあたり、まずは CEFR が設定するレベルの内容を理解・精通するため以下の作業を行う。

1. 調査前、パネリストは課題（Homework Tasks）を与えられる
2. 調査中、CEFR で表されている言語スキルに関する広範囲に渡るディスカッションをし、各レベル別で最低限必要とされる能力を定義する

◎1. 事前課題 （7 ページおよび 36～40 ページ）

CEFR レベル概要・定義を十分理解したうえで、CEFR レベル A1 から C2 それぞれに当てはまる最低限の英語力（○○ができるから○○レベルである）について主要な特徴、指標を個々で記述しておく。考察すべき CEFR 表は、例えばスピーキングに関しては、全体的な口頭表現(Overall Oral Production)、全体的な口頭コミュニケーション(Overall Spoken Interaction)、ネイティブスピーカー話者の理解度(Understanding a Native Speaker Interlocutor)、会話(Conversation)、友人との話合い(Informal Discussion with Friends)。その後、例えば C2 最低レベルにあてはまる受験者にできて C1 トップレベルにあてはまる受験者にできないことは何かなどについて事前に考察しておく。

◎2. 調査中 （8 ページ）
パネリストは、4 技能それぞれの A2、B2、C2 レベルの最低限のスキルを定義し表を作成（事前課題で作成した各自の表と CEFR 表を参照）。
ここで定義するのは各レベルに合致する最低レベル(the candidate who has just enough skills to be at a particular level)の言語使用者についてである。CEFR レベルで表記されているのは各レベルの典型的な言語使用者のレベルなので注意が必要である。その後、全体的なパネルディスカッションを通して A2、B2、C2 各レベルの定義について最終的にパネリスト間で決定する。この A2、B2、C2 レベルの定義を用いて、A1、B1、C1 レベルの最低限のスキルについてパネルディスカッションを通して定義する（パネリストが話し合い最終的に定義したそれぞれのスキルレベルについては Appendix B 46～49 ページを参照）。

基準点の設定
パネリスト全体で各レベルの定義を決定したあとに、実際にスキル別のマッピング作業にはいる。

Reading と Listening セクション（8～10 ページ）
Reading :45 問（Raw Score 0～45） →最終的に 0-30 へ換算
Listening:34 問（Raw Score 0～34） →最終的に 0-30 へ換算

- **Modified Angoff Method** を採用（※問題の設問ごとに、CEFR の各レベルに相当する受験者が 100 人いるとして、正解する者の割合を各パネリストが算出し、全体の平均を合議に基づき判定基準とする分析方法）
- 設問データは、5,000 人以上の受験者が受けているものを使用
- パネリスト（23 人）は事前にカットスコア判断過程について訓練を受け、過程を理解しているかを確認するために実際に判断する機会を与えられる。
- A2, B2, C2 を計 3 回、A1, B1, C1 を 1 回実施

1 回目の判断では、パネリストは、設問ごとに CEFR レベル別の受験者が正答できる確率を 0、5、10、20、30、40、50、60、70、80、90、95、100 の評価基準で出す。確率が高ければ、簡単な問題だと考えられる。この結果、各パネリストが推測するカットスコアが出される。パネルの平均カットスコアと最高、最低のカットスコアが出され、それについてパネリストはディスカッションをし、決定の論拠を共有する。
フィードバックとディスカッションの過程で、Item performance information（P+ values：あるテストフォームの受験者の平均スコア）の指数を共有し、ある設問が受験者の全体的な言語能力に反して比較できるほど難しかったり、異なるレベルの受験者にとって特に難しかったり、簡単だったりするということが分かる。

2 回目の判断をする前に、パネリストは他のパネリストの論拠と基準となる情報について考察する。2 回目の判断では、設問のレベルではなく、セクション全体のレベルについて決定する。2 回目の判断の A2、B2、C2 レベルの平均は、すでに作成されたセクションのスコア分布にあてはめられ、それぞれのレベルに分けられた受験者の割合と照らし合わせて議論する。その後パネリストは、セクションレベルの推奨カットスコアを変更する最後の機会を与えられる（3 回目の判断）。

A2、B2、C2 レベルの最終判断が発表され、その後パネリストは A1、B1、C1 レベルを組み込むよう指示される。

Speaking と Writing セクション（11 ページ）
Speaking（Raw Score 0～24）　→最終的に 0-30 へ換算
Writing（Raw Score 0～10）　→最終的に 0-30 へ換算

- **Modified examinee paper selection method を採用**
  （※受験者の実際の解答を使用して基準点設定を行う）
- Modified Angoff Method と同様、フィードバック、ディスカッション、データの提供などを行いながら、3 回で判断をする。

パネリストは採点基準表（TOEFL iBT Speaking, Writing Scoring Rubrics）を確認し、さまざまなスコアの受験者の 11 の解答を聞くまたは読む。低いスコアから高いスコアの解答の順に並べられている。パネリストには 11 人の受験者の設問ごとのスコアをまとめた表が渡され、各 CEFR レベルの最低カットスコアを判断する。A2, B2, C2 レベルについて 3 回の判断をした後、A1, B1, C1 レベルを組み込む。2 回目以降の判断では、各レベルにとって難しすぎるあるいは簡単すぎると判断した場合は、スコアではなく N/A（該当なし）と回答することも可能。最後にセクションごとに、A1 から C2 の最大 6 レベルのカットスコアを提示する（該当スコアのないレベルも有り）。

---

2. CEFR マッピングを変更するに至った理由、変更点

（参照：”The Association between *TOEFL iBT*® Test Scores and the Common European Framework of Reference (CEFR) Levels”）

理由 1

イギリスを始めとするヨーロッパ諸国では大学入学レベルは CEFR B2 という共通認識があり、それに照らし合わせると TOEFL iBT テストスコアの CEFR との相関表は厳しすぎ、結果とし

て B2 レベルとして表されるレベルを反映するものより高いスコアを求めることになっているのではないかとの示唆がスコア利用者やスコア決定者からあった。（2 ページ）
CEFR を利用している機関から、入学した学生のレベルを見た経験から、例えば B2 レベルのスコア基準を引き下げることは妥当だとのフィードバックがあった。（7 ページ）

理由 2
さらに、ETS アセスメント開発者やスコア利用者が CEFR レベルおよび TOEFL iBT テストが意図する対象言語領域の評価基準の記述(descriptor)についての理解を深めるにつれ、TOEFL テストスコアと CEFR レベルの相関関係について再検証することが妥当であると考えられた。（2 ページ）

CEFR について
特にヨーロッパで広く利用されている言語評価レベル。しかしながら、CEFR はテスト内容の開発やテストスコアの解釈に利用される固定された（評価）ツールでもないし、たった 1 つの正しい解釈や応用のもと従うべき規定でもない。CEFR が広く利用されている理由には、その評価の柔軟性にある。（5 ページ）

スコアの基準設定には批判がつきものだが、これは生来の主観性にある。異なる基準設定方法はいささか異なる結果を生じさせる。同じ受験者が同じテストの別フォームを受験し同じスコアが出るという可能性は低い。したがってテストスコアとスコア基準設定において、ある程度の曖昧さは避けることはできない。（5 ページ）

TOEFL テストスコアの CEFR へのマッピングの修正について（8 ページ）
測定の標準誤差（Standard Error Measurement）
例）1 人の受験者が異なるフォームのテストを受験し同じスコアを取得する確率
68％が 1 SEM（1XSEM）の範囲、95％が 2 SEM（2XSEM）の範囲になる。（詳細は Research Insight Series に記載）ETS はパネリストが推奨した各セクションの最低基準点を再検証し、2SEM の範囲に引き下げることにした。（それぞれのセクションで異なる標準誤差が使用されていたことも留意すべき）その結果、B1 の最低基準点を 57 から 42 に、B2 を 87 から 72 に、C1 を 110 から 95 に変更することになった。
