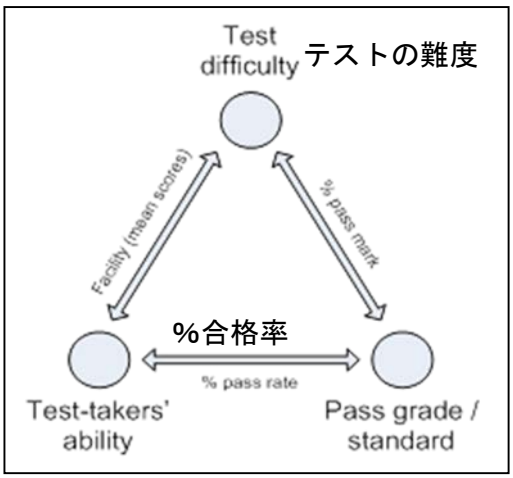
Cambridge Assessment English のアイテムバンキングに対するアプローチ

再現性があり、きちんと説明できる結果を提供するために、評価（assessments）には、誰がいつ評定しても変わらない一貫した基準が必要になる。ケンブリッジ英語検定機構（Cambridge Assessment English）は、項目応答理論（IRT）に基づいたモデルを使用し、出題する問題が同じ尺度上に載るように調整したアイテムバンキングを用意している。

なお、弊機関のプレテスト（事前テスト）の実施および分析は、アンカーリング（複数の異なる時期に実施されたテストを同一尺度上に乗せて相互に比較できるようにすること）や出題する問題の調整を図るためのもので、IRT をテストに適用するために不可欠なものである。

下の図は、テストでの３つの統計概念、容易度（もしくはテストの平均スコア）、合格最低点、および合格率(合格した受検者の割合)を示している。こうした概念は容易に解釈することができる。つまり合格最低点を超えるスコアを取る受検者が増えれば増えるほど合格者数は増える。しかしながら、図が示すように、そのことは単に（３つの中の）２つの基礎的要素の関係を表しているだけであり、それだけでは有益な統計量とはいえない。容易度は受検者の能力とテスト問題の難しさとの関係を表すものである。同様に合格最低点はテスト問題の難しさと合否判定に適用される基準との関係を表している。ここで問題なのは、あるテストでの受検者の総得点（合計点）では、必ずしもその受検者の能力を十分に測ることができないということである。例えば、異なる能力を持つ受検者３名が３つの異なるレベルのテストを受験した場合に、受けたテストで全員が同じスコア（例えば 70%）を取ることも可能性としては考えられる。しかし、こうした測定方式が受検者の真の能力を測るのに役に立たないことは明白であろう。



受験者の能力　　合格グレード/基準

図： テスト実施時の 3 つの基本要素

言い換えれば、容易度、合格最低点および合格率は、それだけでは固有の意味を持たない相対的な概念である。受検者の真の能力、テストのレベル（難しさ）および測定基準の示す意味を知る必要がある。これらは相対的な値ではなく、潜在的な固有の意味を備えた絶対的な値であり、例えば基準についていえば CEFR レベルのような準拠枠（参照枠組）という観点で示すことが可能である。また、受検者は既知の測定結果（レベルの許容範囲）に従い CEFR レベルに照らして、当該レベルよりも下あるいは上のレベルであると特定することができる。

受検者の当該テストでの得点を査定するのではなく、学習者とテストの問題項目との潜在的な関係を見るために、項目応答理論（IRT）をテストの測定評価に用いる。IRT は潜在特性として連続性のある言語習熟度を表すためのアプローチの一つであり、その潜在特性に学習者、能力規準のレベル（基準）をすべて位置づけることができ、互いの関係も定義することができる。テスト応答データから受検者の能力とテストの難易度を引き出す（推定する）には、特定の統計モデル（IRT）を活用することが必要である。弊機関が用いているラッシュ・モデルは IRT に属するモデルの一つである。
(Bond and Fox 2001, Hambleton, Swaminathan and Rogers 1991, Wright and Stone 1979).

ラッシュ・モデルにおける重要な仮定は、受検者がどの問題（タスク）に解答（反応）したかということは重要な問題ではないということである。問題（タスク）の難しさと解答（反応）の内容が分かれば、受検者の能力を十分に測定することができる。他の異なる問題（タスク）でも同一尺度上での能力測定ができていると仮定するには、すべてのテストの問題項目が同一の特性を同一の方法で測定していることが必要である。このことを完全に達成するのは極めて困難ではあるが、開発段階の品質管理には細心の注意を払う必要がある。

テストの素点を用いた尺度とは異なり、ラッシュ・モデルを用いた測定尺度には、私たちが重さや温度のような物理的な性質を測定した時に当然できると考えているものと同じ有用な性質がある。ラッシュ・モデルの測定尺度は、間隔尺度（測定値間の差を比較することに意味を持つ）であり、線形性を持ち、必要に応じて拡張することができる。その性質を備えているため、ラッシュ・モデルを用いてさまざまなテスト事象を関連付けることができ、テストの開発や基準の一貫した適用を一つの枠組みで論じることを促進している。

アイテムバンキング

スコアをどのようにして割り当てるかは、測定しようとする能力によって異なる。リスニングとリーディングは決まった形式で客観的に採点されるが、スピーキングとライティングは専門家により主観的に採点される。同アプローチは計数と評定 (Pollitt,1991)として位置づけられている。評定は、かなり直接的にパフォーマンスを解釈に結びつけるが、計数の場合に

は、項目（アイテム）に対する受検者の応答（正答もしくは誤答）を利用してスコアを直接解釈する前にスコアを変換することが必要になる場合もある。アイテムバンキングは、リスニングとリーディングのように客観的に採点されるテストに用いられる。テストの問題項目は、テストに対する応答を処理するためにラッシュ・モデルを用いて目盛付けされているので、各問題項目の困難度は既に分かっている。そのため、アイテムバンキング（設問のストック）の中から測定したい難易度のレベルに合うような問題を選んでテストを構成し、このようなテストを受検者が受検する。このように、ラッシュ・モデルを用いれば、各受検者の総得点から受検者の能力判定を行うことが可能になる。

最後に例えば CEFR レベルなどが基準として適用される。ラッシュ・モデルを用いることによって、CEFR の基準と当該テストの適切な合格点とすべき難易度との関連付けが可能になる。CEFR 基準は、受検者がテストで達成したグレードを計算するために直接準拠する既知の値で定義されている。同じレベルのテストであるが、わずかに難易度が異なっているような場合でも、CEFR 基準を客観的に適用することができる。

事前テスト、アンカーリング（共通テスト項目の活用）、調整

アイテムバンキングを構築し維持するために、バンク内のすべての問題項目の特性（困難度パラメータなど）を把握する必要がある。その際に項目を特性の値に応じて配置する適切な基準が必要である。問題項目の特性値を推定するためのデータはプレテストによって得ることができる。プレテストには、レベルに合った英語運用能力を持つ学生ボランティアグループに新しい問題を受験してもらうといったテスト業務が含まれる。プレテストによって弊機関では問題項目が適切なレベルであるか、能力が高めの受検者と低めの受検者を十分に識別できているかを確認することができ、さらに予め想定された目的に沿った特性値を項目に付与することができるようになる。

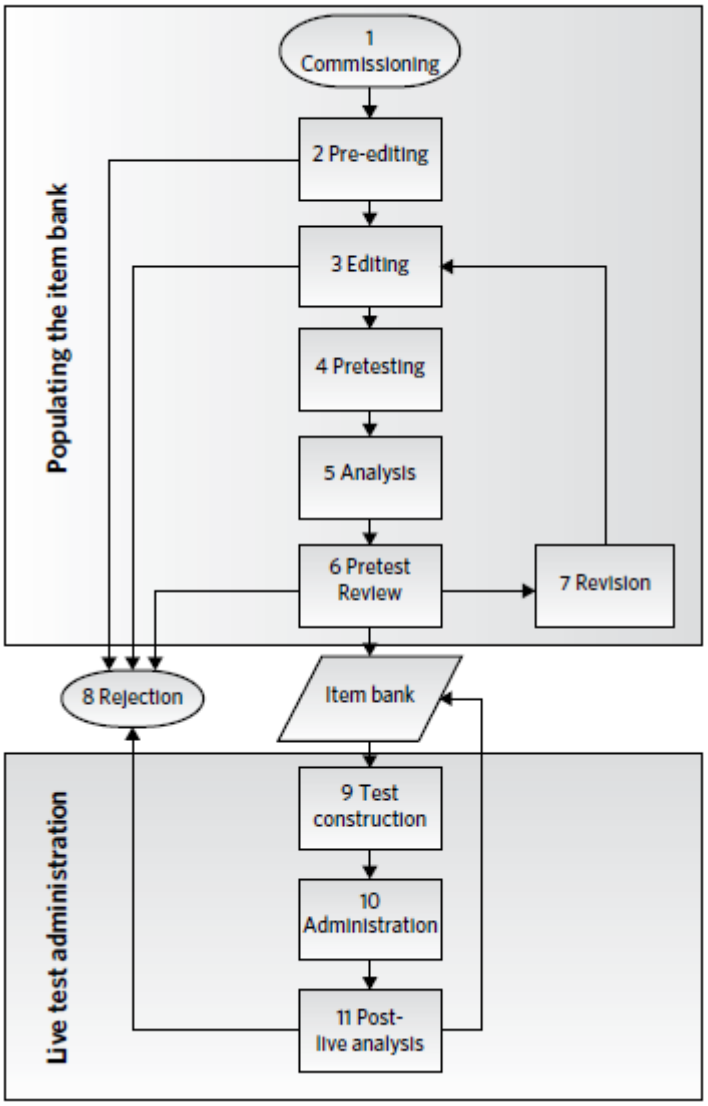
アンカーリングによって、単一の尺度上にデータを位置付けるように目盛付け （calibration）をすることができる。困難度が既に分かっている既存の問題項目（アンカーアイテム）と新しい問題項目とを同時に実施することによって、新しい問題項目の特性をアイテムバンキングと同じ尺度上で表わすことができる。実際には、ラッシュ・モデルを用いて、プレテストの受検者がアンカーアイテムで示した応答（解答）に基づいて、受検者の能力値を推定することができる。

そして、ここで推定された能力値と当該受検者の新しい問題項目に対する解答とを用いて、新しい項目の困難度を推定することができる。

アンカー問題を備えたテストを細心の注意を払って構成し、十分に管理された条件の下でプレテストを実施することによって、弊機関はすべての問題項目を、困難度レベルに対応させて単一の測定尺度上に位置付けることができるのである。

以上のような理由により、適切で一貫した困難度(難易度)レベルのテストを作ることが可能となり、定義された基準を一貫して適用することが可能になる。

Figure 3: The test production and administration process in Cambridge English Language Assessment



Corrigan, M and Crump, P (2015) Item analysis,
Research Notes 59, P9 より抜粋

References

Bond, T G and Fox, C M (2001) *Applying the Rasch Model*, Mahwah: Lawrence Erlbaum Associates.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Hambleton, R K, Swaminathan, H and Rogers, H J (1991) *Fundamentals of Item Response Theory Volume 2*, Newbury Park: Sage.

Jones, N (2000) Background to the validation of the ALTE 'Can-do' project and the revised Common European Framework (PDF 186kb), Research Notes 2, 11–13.
Jones, N (2001) The ALTE Can Do Project and the role of measurement in constructing a proficiency framework. Research Notes 5, 5–8.

Jones, N and Saville, N (2009) European Language Policy: Assessment, Learning and the CEFR, Annual Review of Applied Linguistics, 29, 51–63.

North, B (2008) The CEFR levels and descriptor scales, in Taylor, L and Weir, C (Eds), 21–66.

Pollitt, A (1991) Giving students a sporting chance: assessment by counting and by judging, in Alderson, J C and North, B (Eds) *Language Testing in the 1990s*, London: Macmillan, 46-59.

North, B (2008) The CEFR levels and descriptor scales. In Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity.
Proceedings of the ALTE Berlin Conference, May 2005. Cambridge University Press.

Taylor, L and Jones, N (2006) Cambridge ESOL exams and the Common European Framework of Reference (CEFR), Research Notes 24, 2–5, Cambridge: Cambridge ESOL.

Wright, B D and Stone, M H (1979) *Best Test Design*, Chicago: MESA Press

Jones, N (2014) *Multilingual Frameworks: The Construction and Use of Multilingual Proficiency Frameworks*, Studies in Language Testing Volume 40, Cambridge: UCLES/Cambridge University Press.

Corrigan, M and Crump, P (2015) Item analysis, *Research Notes* 59:4–9.

Council of Europe (2003a) Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.
Manual, Preliminary Pilot Version. Strasbourg: Council of Europe.

Council of Europe (2003b) Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Languages. Cambridge ESOL.