

小泉科研プロジェクト第2回例会 / JLTA第52回研究例会 2021年2月14日

大規模スピーキングテストにおける 採点の運用と課題

井上千尋

英国バッドフォードシャー大学英語学習評価研究所（CRELLA）准教授。
日本の高大で英語を教えたのち、CRELLAでポスドク、講師を経て現職。
JAPAN TIMES ALPHA紙で『英語なるほどQ&A』隔週連載中。

前置き

IELTS、ケンブリッジ英検、TOEFL iBT、Versantに焦点

- どのようなスピーキング能力をどう測っているかに違い
- 公開されている情報量が多い

試験実施にかかる予算や人員など、校内スピーキングテストとは大きく異なるが…

- 測ろうとしている能力、どんなタスクと採点基準が使われているか
- 採点の信頼性がどのように担保されているか
- 制約・課題も多い→校内スピーキングテストの強みは？

内容

1. タスク、採点形式とループリック（採点基準）の紹介
2. 採点の信頼性のために：採点者トレーニング／自動採点の仕組み
3. 大規模スピーキングテストの制約・課題
4. 校内スピーキングテストへの示唆

1-1. IELTSのタスク

- Multi-level (スコア1~9)
- 11~14分
- 試験官と1対1の対面形式
- 試験官が試験の実施および評価を行う



Part	タスク内容	例
1	自身のことや身近な話題についての質問に答える	What's the most interesting part of your town/village?
2	トピックと話すべきポイント3つが書かれたカードを渡され、それに沿ってスピーチを1分間で準備し（メモあり）、2分間で発表する	Describe something you own which is very important to you. You should say...
3	パート2で話したトピックについて、より抽象的なレベルで試験官と話し合う	What kind of things give status to people in your country? Have things changed since your parents' time?

1-1. IELTSの評価

試験終了後、試験官が各観点（4つ）につき1～9点のスコアをつけた後、平均点に基づいて最終スコアが算出される。



評価観点		内容
流暢さと一貫性	繰り返しや訂正の多さ、話せるまとまりのレベル、つなぎ言葉や談話マーカーの使用	
語彙リソース	語彙の多様さ(非頻出語)、正確さ、適切さ、柔軟な使用	
文法の幅と正確さ	文法の多様さ(複雑なもの)、正確さ、適切さ、柔軟な使用	
発音	明瞭度、訛りの強さ、通じやすさ	

公開されているルーブリック(採点基準): <https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en>

1-1. IELTSの評価基準：能力記述の例

流暢さと一貫性（Fluency and Coherence）の一部抜粋

Score 4:

- Cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction

Score 5:

- Usually maintains flow of speech but uses repetition, self correction and/or slow speech to keep going
- Produces simple speech fluently, but more complex communication causes fluency problems

1-1. IELTSのスコアレポート

- 個人情報と顔写真
- 4技能別スコアと全体スコア

観点別のスコアやフィードバックはない

INTERNATIONAL ENGLISH LANGUAGE TESTING SYSTEM
Test Report Form ACADEMIC

NOTE: Assessment is undertaken and your graduate courses should be based on the ACADEMIC Reading and Writing Modules.
CAUTION: The IELTS Reading and Writing Modules are not designed to test the full range of language skills required for academic purposes.
It is recommended that the candidate's language skills are evaluated in the Test Report Form by a specialist after five years from the date of the test.

Centre Number Date Candidate Number

Candidate Details

Family Name First Name Candidate ID

Date of Birth Sex (M/F) Scheme Code

Country or Region of Origin Country of Nationality First Language

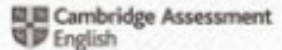
Test Results

Listening Reading Writing Speaking Overall Band Score

Administrator Comments

Centre stamp  Validation stamp 

Administrator's Signature Date Test Report Form Number

The validity of this IELTS Test Report Form can be verified online by recognizing organisations at <http://ielts.org.uk>

1-2. ケンブリッジ英検のタスク



- Key (A2) ~Proficiency (C2) の5レベル
- 8~16分 (レベルによる)
- 受験者はペア、試験官側もペア
- 試験官：進行と全体的評価、評価官：観察と分析的評価

	基本的な構成 (井上による分類。各テストのパート分けとは必ずしも一致しません)
1	試験官が尋ねる、身近な話題についての質問にそれぞれが答える
2✕	それぞれに違う写真／イラストが渡され、それを見ながら相手の受験者に説明する
3	同じ写真／イラストを見ながら、質問を元に相手の受験者とやり取りをする
4	その写真／イラストに関連する話題について、試験官からの質問にそれぞれ答える

✕A2では、2はなし。それ以外は、テストの構成は基本的に同じ。ただし、レベルが上がるにつれて、使われる写真・イラストの複雑さや、話題・質問の難易度が上がっていく。

1-2. ケンブリッジ英検の評価

試験官がつける「全体的達成度」の0～5点評価は最終素点の3分の1を占める。

評価官は、観点それぞれを0～5点で採点し、その合計は最終素点の3分の2を占める。



分析的評価観点		内容
語彙・文法 (B2以上は2観点に)	多様さと正確さ、話せるトピックの抽象度	
談話マネジメント (B1以上)	発話の長さ、まとめ、内容の関連度	
発音	明瞭さ、イントネーションや強勢のコントロール	
相互的コミュニケーション	やり取りの適切さ、会話の開始・保持がどの程度できるか	

A2採点基準:<https://www.cambridgeenglish.org/images/168617-assessing-speaking-performance-at-level-a2.pdf>

B1採点基準:<https://cambridge-exams.ch/sites/default/files/b1-preliminary-handbook-2020.pdf>

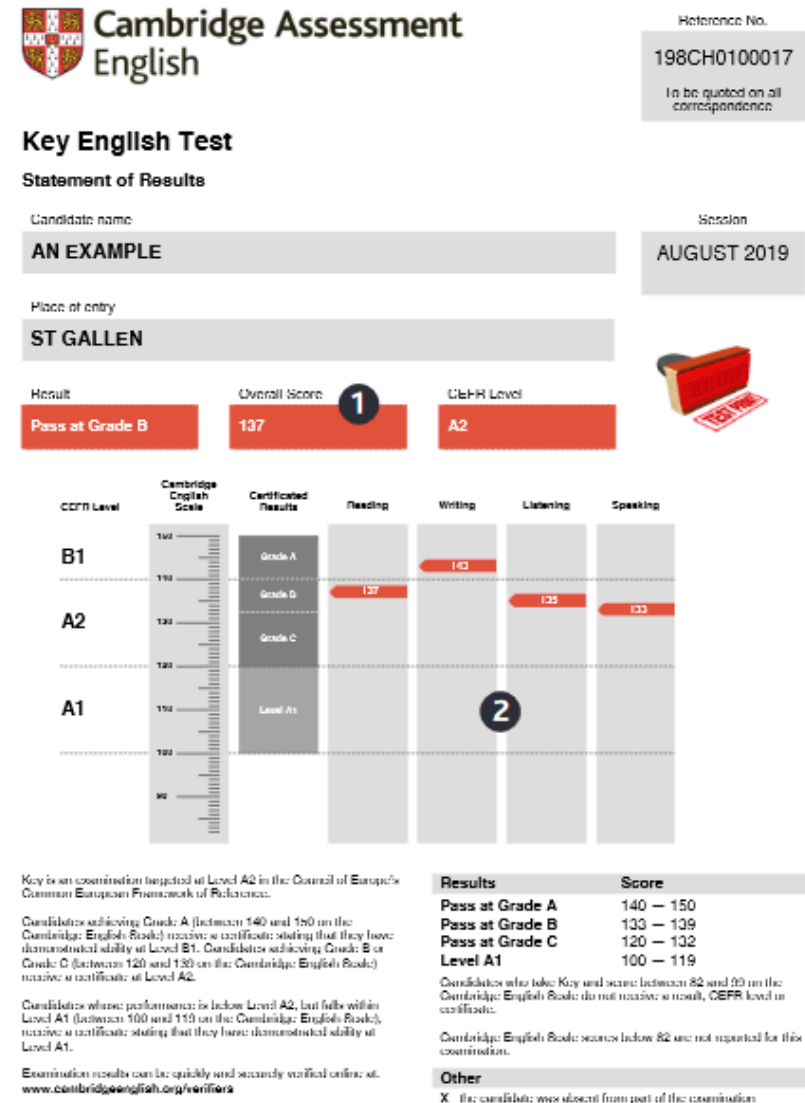
1-2. ケンブリッジ英検の評価基準：能力記述の例

A2	Grammar and Vocabulary	Pronunciation	Interactive Communication
5	- Shows a good degree of control of simple grammatical forms. - Uses a range of appropriate vocabulary when talking about everyday situations.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	- Maintains simple exchanges. - Requires very little prompting and support.
4	<i>Performance shares features of Bands 3 and 5.</i>		
3	- Shows sufficient control of simple grammatical forms. - Uses appropriate vocabulary to talk about everyday situations.	Is mostly intelligible, despite limited control of phonological features.	- Maintains simple exchanges, despite some difficulty. - Requires prompting and support.
2	<i>Performance shares features of Bands 1 and 3.</i>		
1	- Shows only limited control of a few grammatical forms. - Uses a vocabulary of isolated words and phrases.	Has very limited control of phonological features and is often unintelligible.	- Has considerable difficulty maintaining simple exchanges. - Requires additional prompting and support.
0	<i>Performance below Band 1.</i>		

1-2. ケンブリッジ英検のスコアレポート

- 受験者氏名
- 受験場所、日時
- 合否結果・全体スコア・CEFRレベル
- グラフ
- 結果の説明（何点から何点までがいずれのグレードで合格か、など）

観点別スコアやフィードバックはない



1-3. TOEFL iBTのタスク

- 英語圏の大学への留学希望者のためのテスト
- 17分
- PCに向かって話す
- 発話は録音され、自動採点と採点者両方による評価



Task	内容
1	対立する2つの状況あるいは意見が提示され、それに対する自分の意見を述べる
2	大学生活に関連したトピックについての読解テキストと聴解テキストの中の情報を統合し、要約をまとめて話す
3	読解テキストは特定の専門用語を紹介するもので、聴解テキストはその用語について例などを用いてさらに説明する。両方のテキストからの情報をまとめて話す
4	聴解のみを伴う統合タスク。講義の一部を聞き、要点や例を整理してまとめて話す

1-3. TOEFL iBTの評価

- 人工知能（SpeechRater）による自動採点と、人間の採点者1人によるハイブリッド採点方法
- 採点者は発話の内容や意味、言語使用を評価し、SpeechRaterは言語的特徴（発音、訛り、イントネーションなど）を評価する。



評価観点	内容
音声面(デリバリー)	発音、イントネーション、通じやすさ、流暢さ、ペースなど
言語使用	語彙・文法の幅と正確さ、理解を阻害する誤りの頻度
話題発展(内容)	内容の充実度やまとまり具合、タスクに即した回答かどうか、(統合タスクの場合)重要な点が抜けていないか

https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

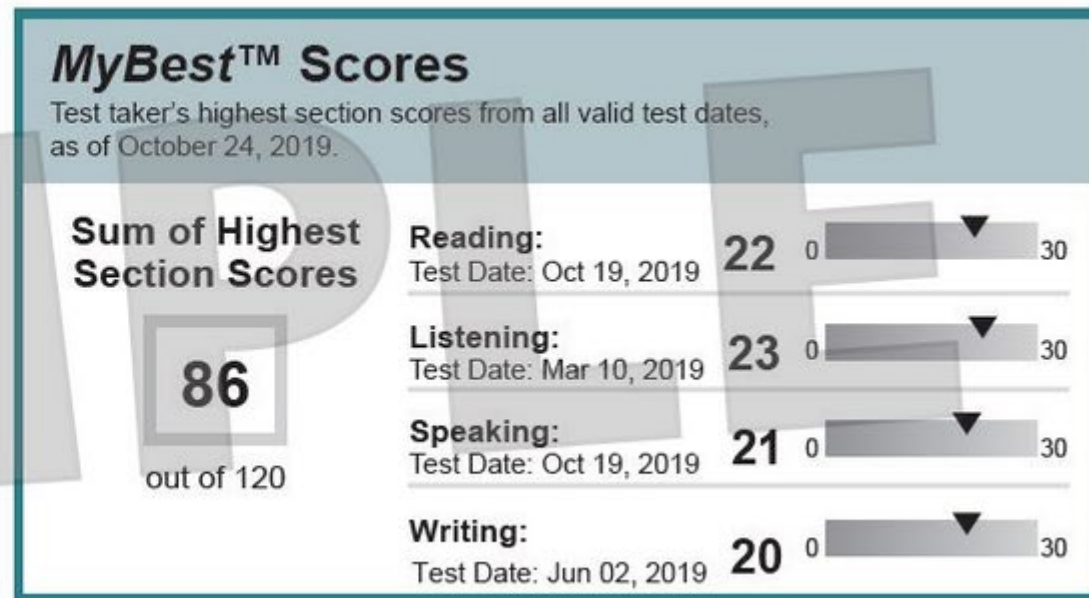
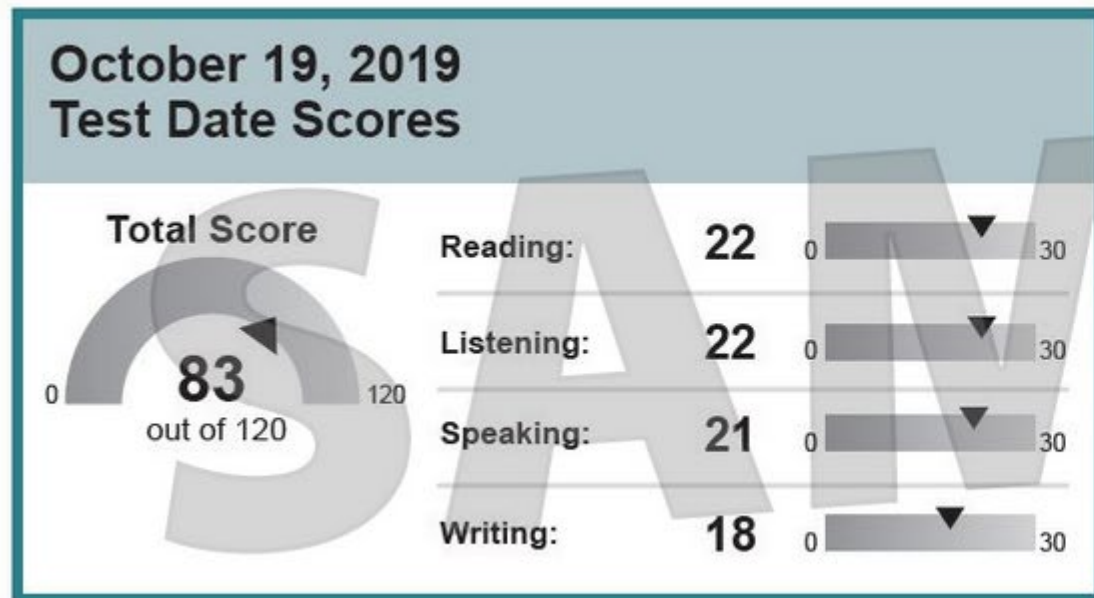
1-3. TOEFL iBTの評価基準：能力記述の例

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.

1-3. TOEFL iBTのスコアレポート

- 個人情報・顔写真
- 全体スコア・技能別スコア・過去二年間のベストスコア
- 結果の説明

観点別スコアやフィードバックはない



1-4. Versantのタスク

- 限定的なタスク
- 文脈に左右されない「言語処理能力」を測る
- スマホ／PCに向かって話す（15分）
- 発話は録音され、自動採点される



Task	内容	例
1	音読	Larry's next door neighbors are awful./ They play loud music all night when he's trying to sleep. [...]
2	リピート	Get some water. / Let's meet again in two weeks.
3	文構築	[she didn't notice / the book / who took] を並べ替え、文にする
4	短答問題	What is frozen water called?
5	要約問題	3～6文（30～90語）の短いストーリーを聞いて、要約を30秒間で話す
6	自由回答	【非採点対象】

1-4. Versantの評価

完全自動採点により、以下の4観点から総合点（20～80点）が算出される。詳細は非公表。



評価観点	内容
文の熟達度	英語のフレーズや節を理解・想起し、完全な文の形で産出する力
語彙	文で使われている頻出単語を理解し、必要に応じて産出する力
流暢さ	文産出、音読、およびリピート時のリズム、フレーズの区切りとタイミングなど
発音	英語の子音、母音、強勢を通じやすく産出する力

妥当性研究レポート：<https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-English-Test-Description-Validation-Report.pdf>

（p.12に、どのタスクの回答がどの観点到に寄与するかの図あり）

1-4. Versantの評価基準とタスクの呼応関係

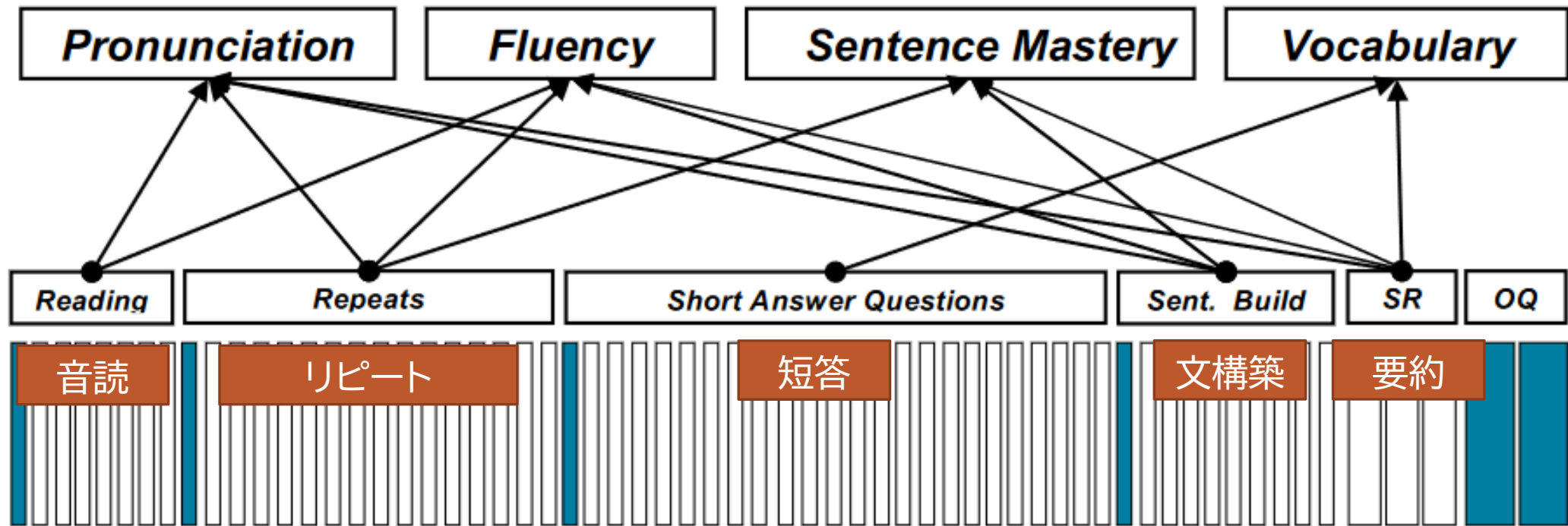


Figure 3. Relation of subscores to item types.

(妥当性研究レポート12ページ目から抜粋)

1-4. Versantのスコアレポート

- テストID番号
- 受験日時
- 受験終了時刻
- 全体スコア・CEFRレベル
- 観点別スコア・グラフ
- 全体スコアに該当する能力記述

具体的なフィードバックはない

SCORE REPORT



Versant Dutch Test

Test Identification Number: 12345678
Test Completion Date: January 1, 2019
Test Completion Time: 1:23 PM (UTC)

SCORE	CEFR
25	B1

SKILL AREA	SCORE	10	15	20	25	30	35	40
Overall Score	25							
Sentence Mastery	28							
Vocabulary	21							
Fluency	26							
Pronunciation	23							

	DESCRIPTION
Overall	The Overall Score of the test represents the ability to understand spoken Dutch and speak it intelligibly at a native-like conversational pace on everyday topics. The Overall score is based on a weighted combination of four diagnostic subscores. Scores are reported in the range from 10 to 40.
Candidate's Capabilities	Candidate can handle many utterances using a variety of words and structures, and can follow and sometimes participate in a native-paced conversation. Pronunciation is mostly intelligible; candidate can express some composite information on familiar topics to a cooperative listener.

© 2019 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s). Other names may be the trademarks of their respective owners.

For more information, visit us online at www.VersantTest.com

PEARSON

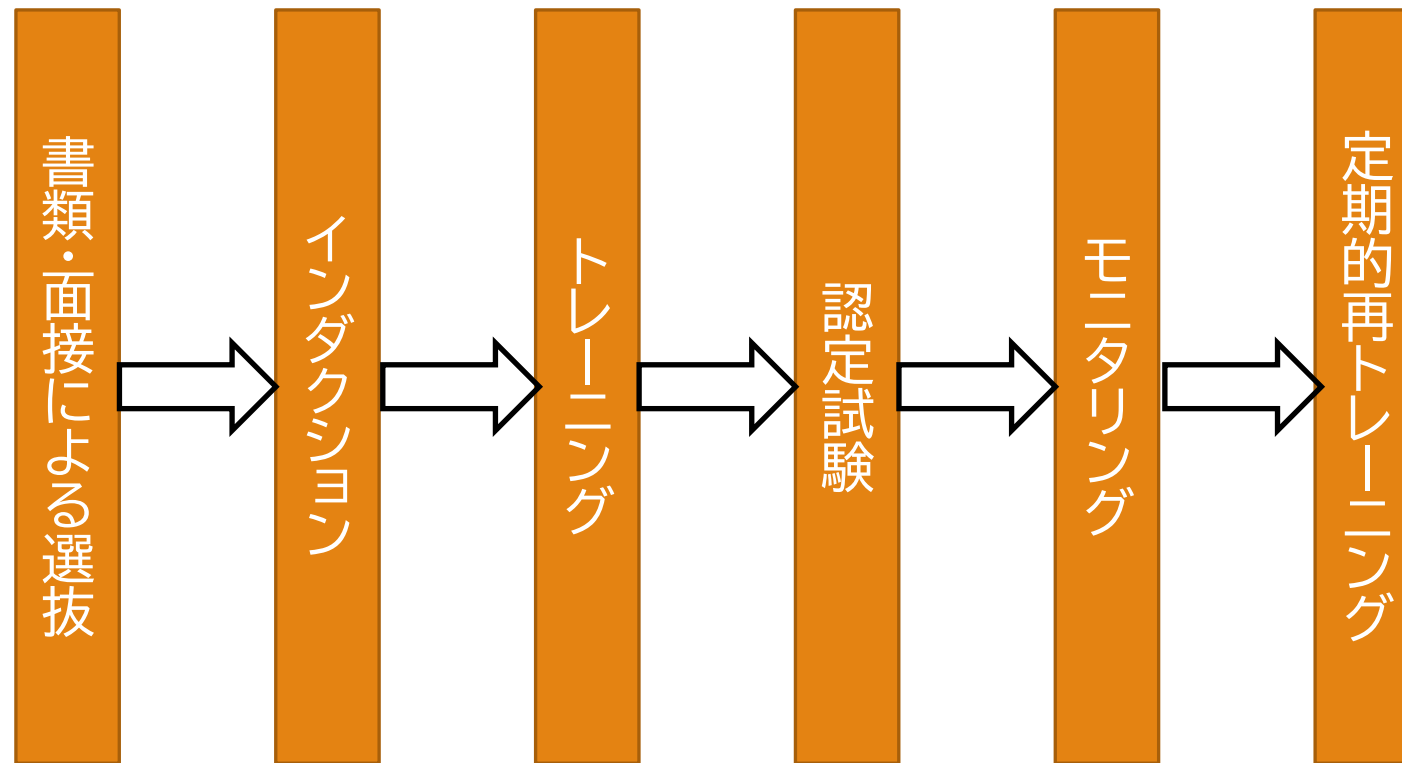


1. 小まとめ

- 大規模スピーキングテストでは、観点別スコアやそれ以上のフィードバックが受験者に返されることは稀
 - ➔ 外部機関に語学力を証明するものなので、そうした需要や期待が低かった
- 観点別スコアやフィードバックがあっても、学習にすぐ役立つかは別
 - ➔ 到達度テストなので、受験者の背景や学習環境が幅広すぎる
 - ➔ 「物差し」が長く、目盛の幅も広い＝どうしてもgeneralなものになる

2-1. 採点の信頼性の担保【人間による採点】

人間が採点を行うテストでは、採点者の募集と選抜、トレーニング、認定およびモニタリングが常時行われている。



- 学士(以上)の学位と、英語教育 (TEFL/TESOL)の資格が必要。教育経験が必須の場合も
- IELTSスピーキングの場合、ライティングと合わせて丸4日トレーニング
- 認定後も、モニタリングや再トレーニングで著しい逸脱がみられれば、認定取り消し

2-1. 採点の信頼性の担保【採点者トレーニング】

- 採点基準にある能力記述があらわすものを、実際の受験者の発話（の録画／録音）を見ながら（聞きながら）理解する
 - ‘occasional minor errors’ ってどんなもの？
 - ‘frequent pauses’ ってどのくらいの頻度と長さ？
- 「こういう特徴を持つ発話は、このレベル」という共通理解を作り上げる

採点基準とトレーニングはセット！

2-1. 採点の信頼性の担保【採点者トレーニング】

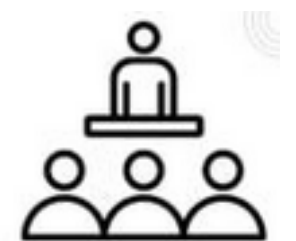
1. ルーブリック（採点基準）の共有・理解

認定トレーナーにより、採点基準と採点の手順が提示され、どのような観点で受験者の発話を評価するのかが共有される。



2. 採点済のスピーキングサンプルの観察

各スコア（あるいはレベル）について、あらかじめ採点されている受験者の発話（録画あるいは録音）が提示され、なぜこの発話とそのスコアなのか、説明がなされる。典型的な受験者の発話にみられる特徴を理解する。



2-1. 採点の信頼性の担保【採点者トレーニング】

3. 集団採点／単独採点

採点済の発話サンプルが提示された後、各自で採点し、その後全員がスコアを開示して話し合う。ルーズブリックを正しく理解して適切に採点できるよう、自分の採点感覚を修正していく。



4. 認定試験

「ルーズブリックを正しく理解し、受験者を適切に評価できるか」が試される。提示されたさまざまな受験者の発話を採点する。認定のためには、正解のスコアと一致していなければならない。



2-1. 採点の信頼性の担保【採点者トレーニング】

校内スピーキングテストでも、複数で採点するなら採点者間信頼性のために

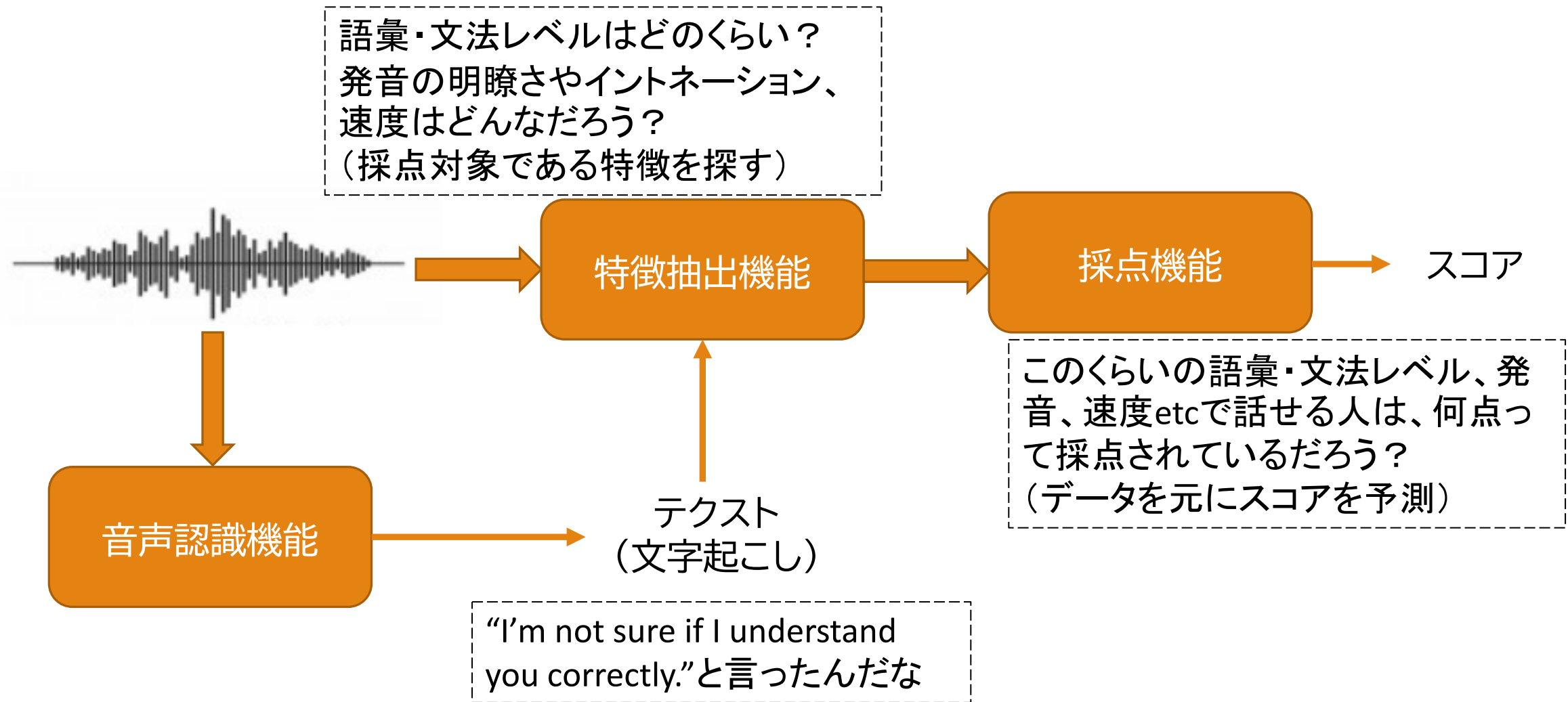
1. ルーブリック共有と理解
2. 採点済の発話サンプルの観察
3. 本番のテスト回答の一部を（最初に）集団採点し、目線合わせ

ができるとうい。

1人で採点を行う場合も、採点者内信頼性を保つために1と2を行い、異なる時点（採点開始時・中盤など）で採点した発話をいくつか再訪し、判断がずれていないかを確認するとよい。

1・2はいつでも参照できるように手元に置いておく！

2-2. 採点の信頼性の担保【自動採点】



2-2. 採点の信頼性の担保【自動採点】

- 機械学習技術によって支えられている
- 「手本」となる発話（採点済み）を大量に学習させた後、受験者の発話がそれと比べてどのくらいのスコアになるかを予測させる
- 機械なので、疲れやブレもない→学習に成功すれば、信頼性は保たれる
- ただし、学習させる「教材」である、「手本となる発話」の採点や選択、および特徴的な指標の選択、学習後の評価や修正には人の手が必要

2-2. 採点の信頼性の担保【自動採点】

- Versant：非常に限定的なタスクを使うことで、受験者の回答の内容や長さのばらつきを小さくし、効率的な機械学習→完全自動採点が可能に
- TOEFL iBT：SpeechRaterが音声面のみに特化しているのは、まだそれ以外の側面の採点の精度が十分でないため
- 「内容」や「言語使用の適切さ」に関する判断は、回答の幅が広いために、機械学習させるための特徴を捉えて絞り込むことが難しい

機械学習が容易な観点(特徴)が優先されている

→測れる能力が限定される

3. 大規模スピーキングテストの制約・課題（1）

- スコアレポートで受験者がもらえる情報が限られている
 - ➔なぜそのスコアになったのかが不明瞭
 - ➔改善のためのステップが把握しづらい
- 受験者を弁別するため、ルーブリックのレベルの幅が広い
 - ➔受験者のレベルが似通っている場合、細かい差を拾えない
 - ➔次のレベルに行くのに大きな進歩が必要

3. 大規模スピーキングテストの制約・課題（2）

- 公平性を期するため、試験官は「台本」から逸脱できない
 - ➔ 行き詰まった受験者を手助けすることは難しい
- 伝統的な「言語能力観」では「話す内容」が重視されてこなかった
 - ➔ 何をもって『良い内容』とするのか？（技能統合タスクのみならず）
 - ➔ 「内容」の評価のばらつきをどう収めていくか

4. 校内スピーキングテストの強み（１）

- スコアレポートで受験者がもらえる情報が限られている
 - ➔なぜそのスコアになったのかが不明瞭
 - ➔改善のためのステップが把握しづらい
- 受験者を弁別するため、ルーズブリックのレベルの幅が広い
 - ➔受験者のレベルが似通っている場合、細かい差を拾えない
 - ➔次のレベルに行くのに大きな進歩が必要

もっと幅の狭い、簡易なルーズブリックを使用し、観点別スコアを返却。
改善のための小さな次ステップを示しやすい
(例:「あと2文足してみよう」「文のどこに強勢があるか注意してみよう」など)

4. 校内スピーキングテストの強み（2）

- 公平性 生徒が行き詰っていたら、手助け(言い換えたり、順を追った回答の引き出し方など)が可能

- 伝統的な「言語能力別」では「話す内容」が重視されてこなかった

より良い内容を持つ回答作成のための指導を経て、テストにつなげる

→ 「内容」の評価のほらつきをどう収めていくか

タスクの種類を柔軟に決められる。学習段階やフォーカスしたい部分に合わせて、音読・リピート・発表・ペア会話・グループ会話など。
採点も、教師だけでなく、生徒同士にさせることもできる

Thank you for listening!



chihiro.inoue@beds.ac.uk



<https://www.beds.ac.uk/crella/>



@crella_beds