教室内技能統合型スピーキングテストにおけるルーブリックと採点

－口頭要約課題を例に－

弘前大学　教育推進機構

横内裕一郎

技能統合型スピーキングテストの特徴

技能統合型のテスト

■4技能を複合的に混ぜ合わせた形式のテスト
e.g., TOEFL iBT
聞く→話す、読む→聞く→書くなど

参考（高等学校学習指導要領（平成30年度告示）p.10
英語コミュニケーションI
…五つの領域別の言語活動及び複数の領域を結び付けた統合的な言語活動を通して，五つの領域の総合的な指導を行う科目である。特に，聞いたり読んだりしたことの概要や要点を目的に応じて捉えたり，基本的な語句や文を使って情報や考え，気持ちなどを話して伝え合うやり取りを続けたり，論理性に注意して話したり書いたりして伝える又は伝え合うことなどができるようになることを目標としている

技能統合型のテスト

利点

■文章を読んだり聞いたりして自身の発表につなげることから、English for academic purposeの観点から真正性が高い
→先行研究を読んで口頭発表につなげるようなイメージ
→授業を聞いてその内容をまとめ、口頭発表する活動のイメージ

■受験者に共通の情報を与えるため、公平性が高い

■受験者が与えられた情報から内容と表現を学ぶことができるため、学習への高い波及効果が期待される（e.g., Barkaoui et al., 2013; Cumming, 2013; Huang, Hung, & Hong, 2016; Plakans, 2007）

技能統合型のテスト

欠点

■与えられた情報を理解する能力が十分でない場合，スピーキングの能力を発揮しにくい
→RやLの能力が低いのか？Sの能力が低いのか判断しづらい。
→どちらかと言えば総合力を問うテスト

（欠点はあるけれど…）

→授業中に触れた語彙や表現を発表させるために問題を調整しやすいので指導と評価を近づけることができる

※しっかりと設計しなければただの暗記テストに変わる可能性

技能統合型のテストは発表？やり取り？

中学校

■社会的な話題に関して聞いたり読んだりしたことについて、考えたことや感じたこと、その理由などを、簡単な語句や文を用いて述べ合うことができるようにする（中学校：やりとり）

■取り入れた情報の内容を話すのではなく、どちらかと言えば自身の考えを述べさせる課題が求められる

技能統合型のテストは発表？やり取り？

高等学校

■社会的な話題について、使用する語句や文、発話例が十分に示されたり、準備のための多くの時間が確保されたりする状況で、対話や説明などを聞いたり読んだりして、情報や考え、気持ちなどを理由や根拠とともに話して伝える活動。また、発表した内容について、質疑応答をしたり、意見や感想を伝え合ったりする活動（高等学校：発表）

e.g., summary speech, retelling

■与えられた情報の内容を理解し、他者に伝える能力が求められる

➔技能統合型のタスクはやりとり・発表どちらにも応用可能

技能統合型問題 vs 独立型問題

	技能統合型問題	**独立型問題**
妥当性	△（LやRの影響を受ける可能性）	◎（Sを直接測定できる）
公平性	◎（情報の統制が可能）	△（背景情報が発話に影響する可能性）
真正性	○（実世界でありうるタスク）	○（実世界でありうるタスク）
波及効果	◎（学ばせたい表現を使用することを強制可能→学習効果大）	○（avoidanceが多発する可能性）
問題作成のしやすさ	○（刺激文を作成・調整する必要性）	○（問題形式によって問題作成のしやすさが大きく変わる）
評価のしやすさ	○（回答がある程度制限される）	△（回答が多岐にわたる可能性）

小泉, 印南, 深澤. (2017)、Luoma (2004)を参考に作成

技能統合型テストの作り方

■タスクの種類（一例）
・要約
・再話
・再生（暗記になりがち）
・絵を使った統合型タスク　等
■情報の与え方
・リーディング形式で与える
・リスニング形式で与える
・両方（information gapを設定する必要あり）

技能統合型スピーキングテストの実践

－口頭要約課題を例に－

要約課題とは

■要約課題とは、比較的長い文を読み、重要な情報（Topic sentence等）を見つけ明確化し、不要な情報を削除したり、必要な情報を選択したり、表現を一般化したり、内容の再構成などを行って、自身の言葉に置き換える活動（Kissner, 2006; 卯城他, 2012, p. 73-74）

➔比較的長い文章を読んだり聞いたりして得た情報を自身の言葉に短く置き換えて発話する（あるいは文章を書く）活動

■話す時間を短めに設定する（刺激文の情報量に応じて変化：300語程度であれば2分程度が妥当）

情報の考え方

リーディングの場合

■暗記出来ない程度の長さの文章を用いる

■文章の難易度は抑えめにする（読解できないと話せない）

■注釈はできるだけつけない
→注釈にある単語を繰り返し使うようになり、表現の幅や内容が偏る要因になりうる
→ターゲット表現を設けて特定の表現・文法を話させたい場合には注釈は有効

■1分あたり何語読むかを検討した上で刺激文の長さを決める
→読む時間もそれに従って決める（高校生だと1分60語程度？）

刺激文の考え方

リスニングの場合（リーディングで検討する事項に加えて…）

■再生速度を考える

・事前に音声を録音する場合はメトロノームなどを使い速度を固定

・合成音声を使用することも検討

→Google TTSなどはかなりの精度

→MacOSやiOS、Windowsの読み上げも有効

→いずれも外部出力で外部機器に録音するのが容易

→デスクトップの音声を録音するのも可

■再生回数を検討する（聞く回数は1回のみ？複数回？）

刺激提示→発表までの準備時間

■準備時間の設定
- リーディングの場合読む時間＝準備時間と考える
- 刺激文を読ませた後、聞かせた後に考える時間をとるかどうか
- 1分程度準備時間をとると、発話が増える傾向がある

(Yokouchi, 2015)

■準備中の筆記
- 準備中に筆記を許すとそれを朗読するだけの生徒が現れる

→音読はスピーキング能力を測定するタスクとして不適当

- メモをとっても良いが、話す際には見ないよう指示を徹底する

要約課題で測るべき能力

■何を測定したいのか・評価したいのかで考える
・流暢さ
・文法・表現・語彙・発音等の正確さ
・文法・表現・語彙等の複雑さ
・発話内容の正確さ（発話の一貫性・刺激文の内容理解度）

→実際に教室レベルで採点できるのは多くて3,4要素程度

■分析的評価は全体的評価より高い信頼性が得やすい（Hamp-Lyons, 1991）
→全体的評価は信頼性の問題に加え、フィードバックが具体的にならないのであまりおすすめしない

測定しやすい能力/ しづらい能力

■流暢さ：音声を聞いただけでだいたい判断できる
　　書き起こしがあれば一目瞭然

■発音　：開始数秒で評価ができてしまう場合も多い

■文法・表現・語彙等の正確さ：

　じっくりと録音を何度も聞く、あるいは書き起こしを観察する必要があるため時間がかかる

■文法・表現・語彙等の複雑さ：同上
　フィードバックしづらい

測定しやすい能力/ しづらい能力

■では、発話の内容面は…？
比較的評価が安定しづらい傾向（Yokouchi, 2018; 2020）

■発話内容の正確さ（発話の一貫性・刺激文の内容理解度）
・刺激文の内容を適切に話しているか
・話の内容に一貫性があるか
・（意見感想を述べるタスクを追加している場合）
→意見感想が刺激文の主題に沿っているか　など

評価観点（構成概念）

■何を測りたいのかを明確にする必要があるが、特に要約活動では発話の内容が重要な評価観点（構成概念）となる

■しかし、どのような観点で要約の内容を評価するべきかは評価者間で考え方に大きな差があると考えられる
→内容の評価基準は特に明確に示す必要がある

ルーブリック（評価基準）

■測るべき能力（構成概念）を決めたら、それぞれの能力でどのようなパフォーマンスを発揮していれば何点を与えるかを決め、ルーブリックを作成する

■必ずしもルーブリックは表の形式になっている必要はない
→教員にとって採点しやすく、学習者（受験者）にとって自分の長所と短所がわかりやすいものが望ましい

■既存のルーブリックを流用する際は、そのルーブリックが同じタスクを使用しており、同じ（あるいは近い）レベルの学習者を対象としたルーブリックを使用する

ルーブリック（案）

	流暢さ（発話量）	発音	文法	内容
5	100語以上の発話	発話に強弱がある 発音ミスが少ない	聞き手が混乱するような表現がない/ 少ない	与えられた文章の必要な情報をすべて話しており、話の内容に一貫性がある
3	50語以上の発話	発話に強弱がない 発音ミスが少ない	内容を推測する必要のある表現が多少ある	与えられた情報の内容を部分的に話している/ [5]点の条件を満たしているが一貫性に欠ける
1	49語未満の発話	発話に強弱がない 発音ミスが多い	内容を理解することが難しい表現が多い	与えられた情報の内容をほとんど話せていないが関連する内容を話している
0	20語未満の発話			

流暢さ→filled pauseやrepetitionが多い場合上記からダウングレード
発音→流暢さが低い場合リズムに影響がでることがあるため観点はイントネーションと発音
文法→小さな誤りはカウントしない
※「ミスが少ない」などの表現は学習者には具体的に何割程度ミスがあると原点か伝えると良い

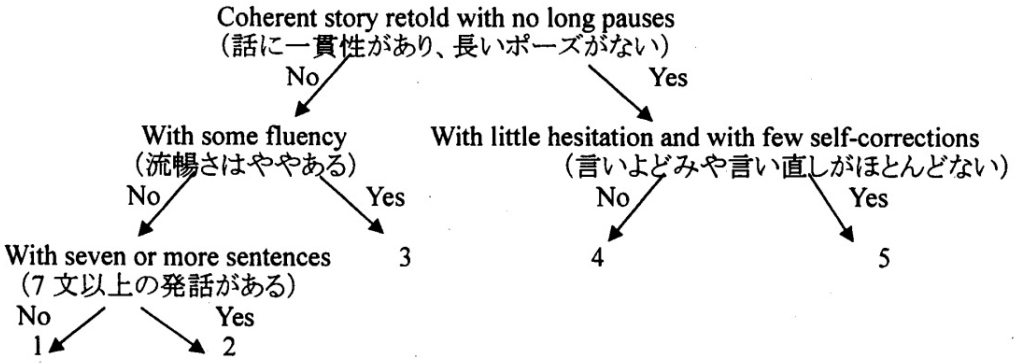
要約課題における発話内容の評価の難しさ

Yokouchi (2018)より

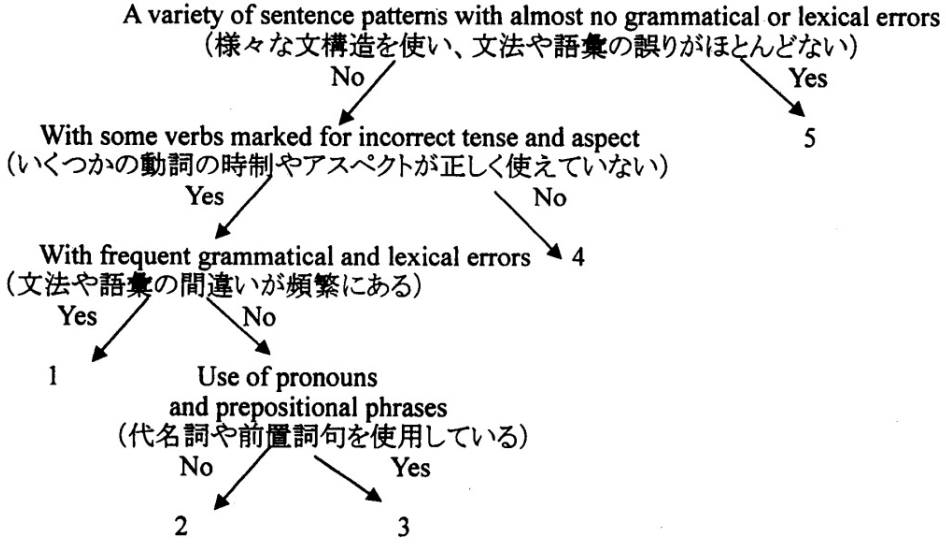
- Hirai and Koizumi (2008) の4項目のEBB（Communicative efficiency, Grammar & Vocabulary, Contents, Pronunciation: 1-5点で評価)に0を付け加えたEBB scale＋全体的評価尺度 (Yokouchi, 2018) を使用
- 128名のスピーキングパフォーマンスを4名の評価者が評価
- 各項目の評価の有効性の観点からどの評価基準が機能したか、機能しなかったかを多相ラッシュモデルによる分析で確認

※Hirai and Koizumi (2013) でEBB scale は更新されて、Contentsの基準が削除されたが、本研究では内容面を重視するためにあえて2008年版のEBB scaleを使用した

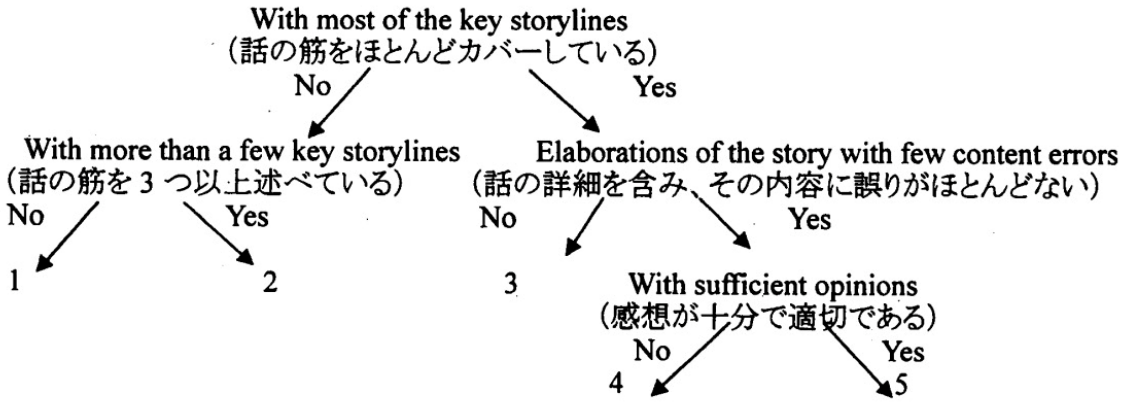
1. Communicative Efficiency（伝達能力）



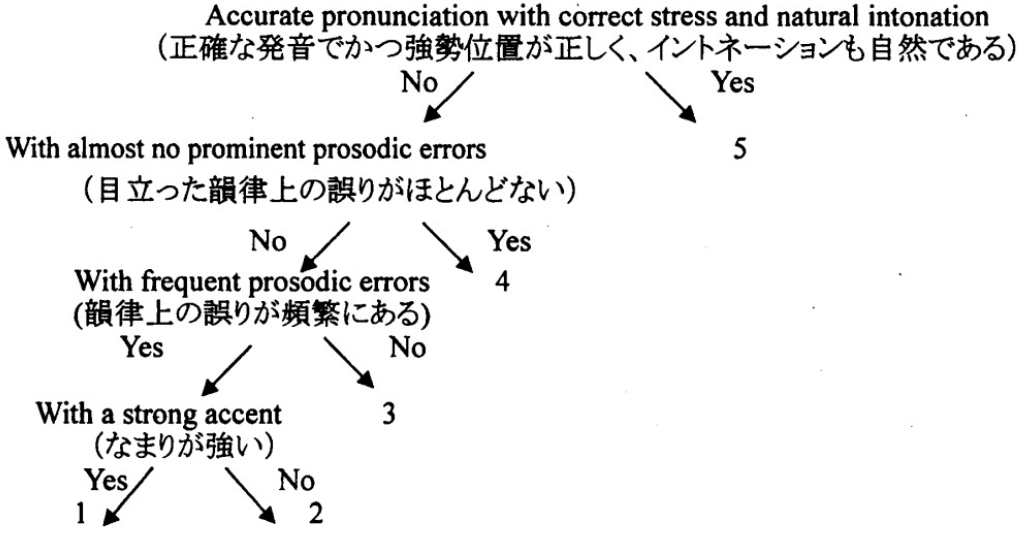
2. Grammar & Vocabulary（文法と語彙）



3. Content（内容）



4. Pronunciation（発音）



Hirai and Koizumi (2008, pp. 19–20)

要約タスク例

以下の文章を5分で読みなさい。その後、読んだ内容を要約して英語で話し、文章の主題に対する意見・感想を述べなさい。解答時間は最大2分です。

Devices such as smart phones, tablet PCs, and laptops are used around the world. People can use them anytime, anywhere to find information or to communicate with others. Many people benefit from them, but some people, especially older people, say that those items are poison.

（以下略、全体で301 words）

回答例1

Devices ah such as smart phone, PC and laptop are used used round the world and in modern world.

So, digital, these devices is ah the only one only one machines in modern world.

Ah, so, so

But, ah especially smart phones, while while people use using smart phone, so they can get into traffic accident.

And, ah crimi criminal… criminals use devices, too.

So…

We need to we need to we need to know and get to to knowledge knowledge getting …of knowledge of application.

評価の例（Hirai & Koizumi, 2008)を用いて

※以下の点数は4名の評価者による評価の結果

1. Communicative Efficiency: 1, 2, 1, 1
2. Grammar & Vocabulary: 1, 1, 1, 1
3. Contents: 1, 1, 1, 1
4. Pronunciation: 2, 2, 2, 4

コメント：発話の大筋を捉えておらず、filled pauseやfalse startが頻繁にある。主に刺激文中の固有名詞と基本的な動詞のみを使った発話となっており、発音に抑揚がない。Rの能力向上と発表語彙を十分に身につけることが上達の近道だろう。

回答例2

Smartphones and the things like that are so good and nice, and be used um for so many people and all over the world.

If if you have that, you could do anything.

And, it's ah so nice, but sometimes especially for older people, that might be poison.

Um, Now business person or students or the thing like people use that to make documents or to ah make repots.

That is so useful.

And if you have smartphones and the things like that, you could do anything ah as a voice recorder or…like that.

(以下略, フィラー込で171 words)

評価の例（Hirai & Koizumi, 2008)を用いて

※以下の点数は4名の評価者による評価の結果

1. Communicative Efficiency: 5, 5, 5, 5
2. Grammar & Vocabulary: 5, 5, 4, 5
3. Contents: 4, 4, 4, 4
4. Pronunciation: 4, 5, 4, 5

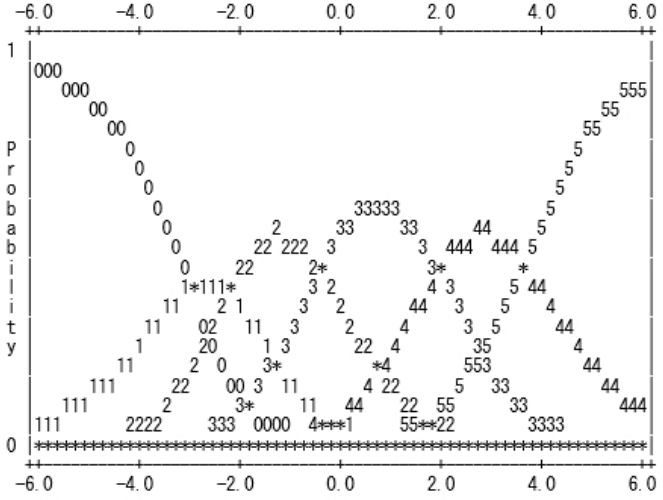
コメント：内容面はすべての情報を網羅しているが、意見・感想がなかったため評価が下がった。文法・語彙も十分に扱えており発音も大きな癖がない。ただし、時折口ごもった発音が見受けられる。口（舌や唇）の動かし方にも注意を払えるように訓練しよう。代名詞の用法にも注意しよう。あとは things likeを多用する傾向があるのでそれを減らせればなお良し。

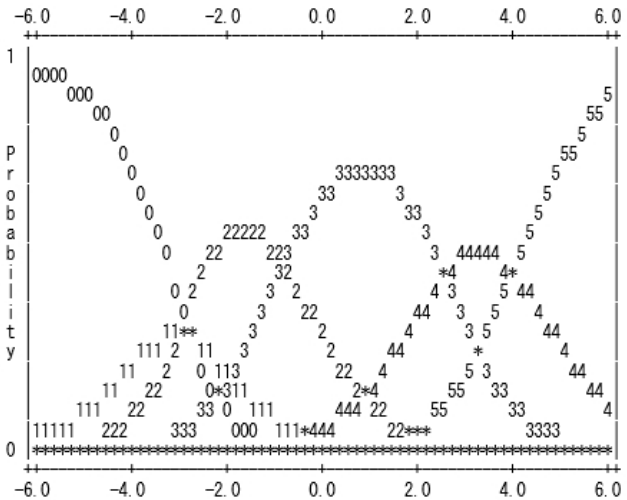
評価基準の適切さの調査 (Yokouchi, 2018)

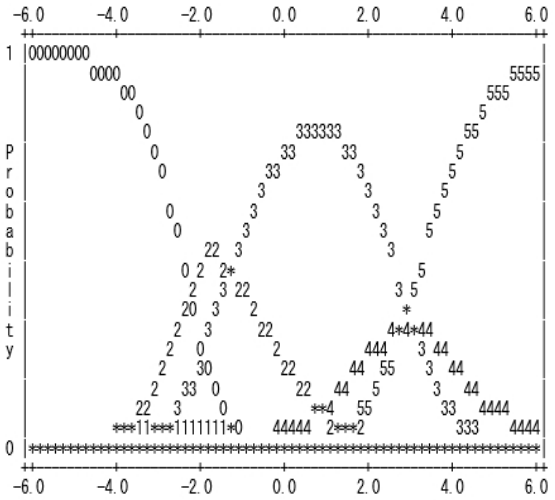
Communicative efficiency
→Probability curvesを見ると、0～5の全てにピークがある（“１”だけは小さい）

Grammar & Vocabulary
→1がほぼ機能していない

Contents
→全体的に機能していない（1,4点はデルタの逆転）
1点は完全に機能していない

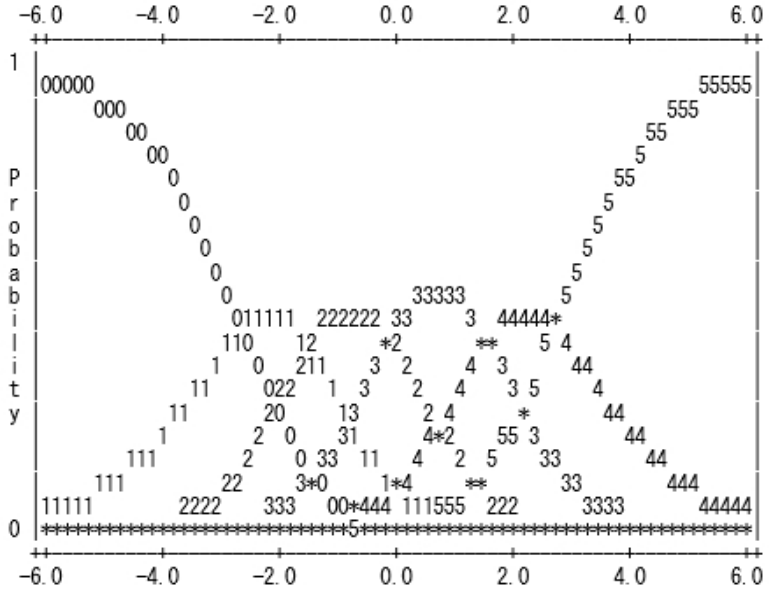




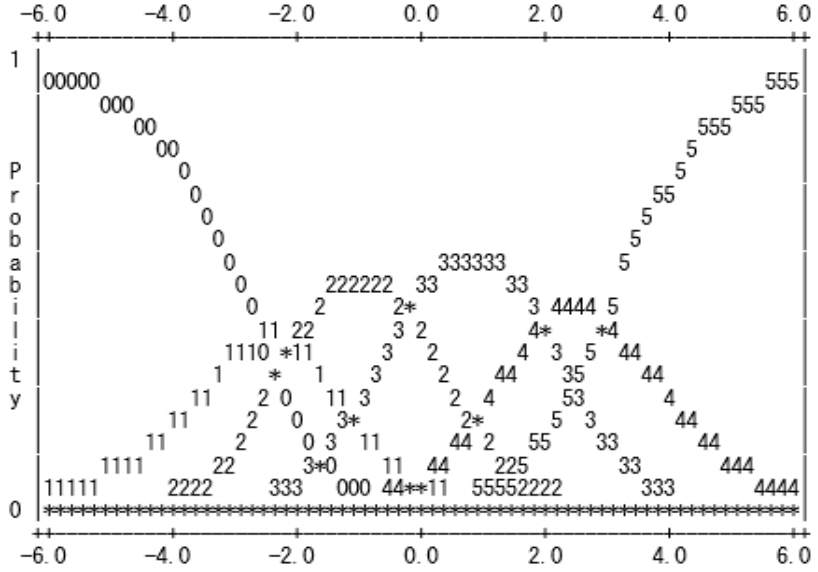


CriteriaのProbability curves

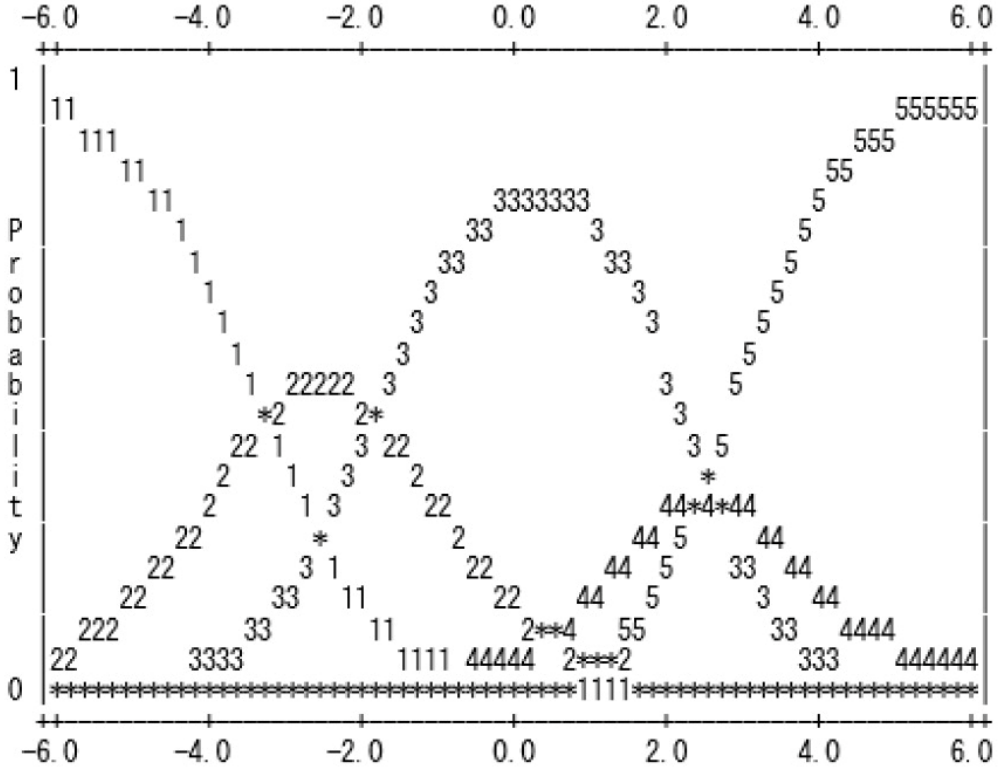
Pronunciation
→5つの評価基準の中では一番綺麗に別れており、評価が適切であったことがわかる



Holistic
→1が機能していない



Hirai and Koizumi (2008) をそのまま使って評価した場合（Contents）



先程の例よりはマシだが、1・3・5点の比率が高く、4点の部分はデルタの逆転が起こっており、4点の評価基準が正しく機能していない。また、2点も1点・3点・5点に比べて該当するlogitの範囲が極端に狭い。

※logitの値は1.4〜5.0離れていればよい（Bond & Fox, 2007）ので、1.71と運用上ギリギリの範囲だと判断できる。

■上記の結果から、Contentsについては評価がうまく機能していなかった可能性が極めて高い

■受験者は同じ文章を読んで発話しているのに（同じ刺激が与えられるのは要約をはじめとする技能統合型タスクの利点）、内容の評価のばらつきが大きかった
→評価者の問題？ or 受験者の能力の問題？

→内容の評価基準設計＆実際の評価時には注意が必要

評価時の工夫

評価時の工夫

書き起こしを活用する

学習者自身に発話を書き起こさせる
→可能であればフィラー（ah, um, wellなど）も書き起こす
→学習者自身に書き起こしをベースに誤っていたと思われる箇所に修正を入れさせる（コメント機能を利用）
→フィードバックしやすい

クラスメイトに書き起こしをさせることも有効

→他の生徒に聞かれることを嫌がる生徒も多い

→他の生徒の発話を聞くことは学習につながる

スクリプトありの場合でも採点時には絶対に音声はすべて聞くこと

（書き出された英語と発話された音声には大きな隔たりが）

評価時の工夫

■ルーブリック・刺激文は常に見返しながら評価する

■マテリアルは何度も読んだり聞いたりできるように用意しておく（学習者の発話を繰り返し聞くことで評価者に悪影響が…）
→発音や文法は特に基準があいまいになりやすい
→必要に応じて評価済みの他の回答を聞き直すことも有効
→やりすぎると負担が大きくなるので注意

■自分なりの満点の回答を録音しておき（できれば４点、３点などの学習者の典型的なパフォーマンスも保存して聞き返せると良い）、いつでも聞き返せるようにしておく（ネイティブに協力してもらっても良い）

書き起こしを使った評価の有効性

■受験者（学習者）に自身の発話を書き起こさせる活動は評価の信頼性を向上させる意味で有効

■2018年度（書き起こしデータを確認した際の評価者内信頼性）
（全体的評価） $\kappa = .91$

■2019年度
（書き起こしデータを確認しなかった際の評価者内信頼性）
（全体的評価） $\kappa = .72$

■注. 要約タスクの採点に慣れている自分でも書き起こしデータを確認しなかった場合には信頼性が低下

まとめ

■口頭要約課題を使用する場合は、受験者のR/Lの能力が一定水準以上あることが前提
→最低でも高校生以上に課すべき課題

■評価基準は項目数を限定して、内容面にも踏み込んで評価する
→内容の評価は難しいので評価基準の設計をしっかりする
→評価者訓練をしっかり行い、安定した評価ができるようにする

■受験者自身に書き起こしをさせる活動は有効
→書き起こしデータに依存せず、音声データはしっかり聞く

参考文献

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics*, *34*, 304–324. https://doi.org/10.1093/applin/ams046

Cumming, A. (2013). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 216–229). John Wiley & Sons Ltd.

Hirai, A., & Koizumi, R. (2008). Validation of the EBB scale: A Case of the Story Retelling Speaking Test. *JLTA Journal Kiyo*, *11*, 1-20. https://doi.org/10.20622/jltaj.11.0_1

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a Story Retelling Speaking Test. *Language Assessment Quarterly, 10,* 398–422. https://doi.org/10.1080/15434303.2013.824973

Kissner, E. (2006). *Summarizing, paraphrasing, and retelling: Skills for better reading, writing, and test taking*. Heinemann.

小泉利恵・印南洋・深澤真 (編著) (2017). 『実例でわかる英語テスト作成ガイド』 大修館書店

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons, L., （Ed.). Assessing second language writing in academic contexts. Ablex, 241-276.

卯城祐司他 (2012) 『英語リーディングテストの考え方と作り方』 研究社

Yokouchi, Y. (2015). The effects of a reading aloud activity as a pre-task on the performance of oral retelling tasks, *TELES (The Tohoku English Language Education Society) Journal, 35*, 93–103.

Yokouchi, Y. (2018). *Effects of task conditions on spoken performance in retelling*. (Unpublished doctoral dissertation). University of Tsukuba, Japan.

Yokouchi, Y. (2020). Correlation between EBB scale scores and CAF indices : Evidence from speakers' actual performances. *Learner Corpus Studies in Asia and the World, 5*, 67-78.