多相ラッシュ分析を用いた GTEC CBT スピーキングにおける評価者信頼性の検討

小泉　利恵（順天堂大学）・岡部　康子（進学基準研究機構：Center for Entrance Examination Standardization: CEES）・鹿島田　優子（（株）ベネッセコーポレーション）

キーワード：スピーキング評価，評価者間信頼性，評価者内信頼性

１．はじめに

大学入試において、スピーキングを含めた 4 技能が測られる方向で議論が進んでいる。学習指導要領で 4 技能の伸長を求め、社会でも、相手とやりとりをしながら、情報の伝達やそれに対する意見が求められる中、その力を伸ばし、評価をし、その結果に基づき、さらに力を伸ばすために指導や学習を改善するという流れが求められている。

スピーキング能力の評価は、指導内容や現実社会でのタスクとの一致、また波及効果の点等から重要であるが、テストを受けた後に得られるテスト結果や、その使用法が適切でなければ、テストの有用性は限られる。そのためテスト機関は、テスト得点の解釈や使用法が十分に適切かを調べる妥当性検証を行い、テスト内容やテストの質、採点の手順などのテスト運営と管理についての詳細を、テスト使用者に公表する責任を伴う（American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Japan Language Testing Association [JLTA] Steering Committee & JLTA Language Testing Terminology Committee, 2006）。妥当性には、専門家による内容的な検討、テストが測る能力が意図したものと一致しているか等、様々な観点があり、その中に、得点の一貫性（信頼性）が含まれる（Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Messick, 1996）。

スピーキング評価には評価者による判断が伴い、評価者が一貫した得点を出せないと、評価者の信頼性が低くなる可能性がある。大学入試に使用される影響力の強いテストでは、高い信頼性を保つことが重要であり、それを保つ手順の確立と実施後の検討が必要である。本研究では、Global Test of English Communication Computer Based Testing (GTEC CBT) のスピーキングセクション（SS）における評価者の信頼性を、多相ラッシュモデルを用いて吟味する。

多相ラッシュ分析を用いた大規模テストの研究は海外では多い（例：Eckes, 2005; Hagan, Pill, & Zhang, 2016）が、日本での使用は限られる（使用した例：Akiyama, 2003; Negishi, 2015; Koizumi, In’nami, & Fukazawa, 2016）。多相ラッシュ分析では、受験者とタスク以外に、評価者等、3 つ以上の相を入れることができる。この分析を用いなくても評価者信頼性を検討することは可能であるが、多相ラッシュ分析により、素点を使った分析では得られない、詳細な情報を得ることができ、有用である。

２．研究課題

GTEC CBT のスピーキング評価について、以下 4 点を検討する。

(1) 評価者の厳しさに違いはあるか（評価者の厳しさの値（Measure）を用いて検討）。
(2) 評価者間での一致度は高いか（評価者間の一致度の％を用いて検討）。
(3) 評価者内の一貫性は高いか（平均平方（Infit mean square）を用いて検討）。
(4) 評価者と評価観点、評価者と受験者の間に偏った評価傾向は見られるか（バイアス分析を用いて検討）。

３．方法

(1) 使用データ

受験者 648 名、SS の 23 観点、評価者訓練を受けた評価者 13 名の採点データを用いた。2015 年に実施した SS データの一部であった。受験者は高校 1～3 年生が中心だった（70％）。

評価者は、GTEC CBT SS を評価した経験が 2 年以上あった。全ての評価者は大学で英語を

専攻した経験があった。

(2) GTEC CBT SS の内容・方法

コンピュータ画面にタスクが出され、それに基づいて受験者が話し、その録音を評価者が後で採点する形式で、約 20 分のテストである。タスクでは、学校生活で英語を使う状況で、やりとりと発表を引き出す（Benesse Corporation, 2016）。パートは 3 つあり、(a) 会話応答問題（小問 6 問）、(b) 情報伝達および照会問題（3 問）、(c) 意見展開問題（3 問）である。1 つの問には複数のタスクが含まれ、タスクごとに評価規準が 1〜5 観点設定されていた（例：質問の意図に沿って、適切な応答ができているか。流暢さ）。計 23 観点を独立して分析した。各タスクの満点は 1 点のものが 10 観点、2 点のものが 11 観点、3 点のものが 2 観点あった。

採点の際には、評価者に対して評価訓練を行い、評価成績の良い者のみが評価者に採用される。実際の評価では、1 観点につき、通常 2 名で独立して評価を行う。評価が一致しなかった場合には、第 3 の経験豊富な評価者が最終的な得点を独立して決める。

(3) 分析

Facets (Version 3.71.4; Linacre, 2014) の部分クレジットモデル（partial credit model）を用いて多相ラッシュ分析を行った（多相ラッシュ分析の詳細は Barkaoui, 2013; Eckes, 2011 を参照）。受験者、タスク、評価者の 3 相を含めた。全体的に SS の評価について調べるために、第 3 の評価者の得点も含めた。なお、第 3 の評価者の得点も含めた場合と、第 1・2 の評価者の得点のみ含めた場合での結果はほぼ同じだった。

本研究は、大学入試に関わる重要度が高いテストが対象のため、Eckes (2005) に基づいてラッシュモデルへの適合（フィット）の基準を決めた。具体的には、インフィット平均平方とアウトフィット平均平方の2つを用い、その判断基準も2つ用いた。第一の基準は、広くてより甘い基準で、平均平方が0.50〜1.50内ならば、ラッシュモデルが予測する一般的なパターンに沿って、モデルに適合したと考えるものである（Linacre, 2013）。0.50未満であればモデルの予測に一致しすぎているオーバーフィット（過剰適合）、1.50を超えればモデルの予測に一致していないアンダーフィット（またはミスフィット）と解釈した。第二の基準は、狭くより厳しい基準で、平均平方が0.70〜1.30内ならば、ラッシュモデルが予測する一般的なパターンに沿って、モデルに適合したと考えるものである。0.70未満であればモデルの予測に一致しすぎているオーバーフィットと、1.30を超えればモデルの予測に一致していないアンダーフィットと解釈した。アンダーフィットは一般的でないパターンであるため問題であるが、オーバーフィットは予測に一致しすぎているという意味で、あまり問題視されないことが多い。

４．結果と考察

図 1 は結果の全体像を示す変数マップである。0 が平均であるロジット尺度上で結果が示されている。値（Measure）が高くなるにつれて、受験者の能力が高く、評価観点が難しく、評価者の評価が厳しくなると解釈できる。結果を数値で表したのが表 1 である。受験者の層（Strata）は約 6 で、今回の SS を用いると、受験者をスピーキング能力で 6 グループに分けることができていた。評価観点の層は約 30 で、難易度の観点から、評価観点が 30 グループに分けられていた。

表 1　3 相の記述統計

	平均値 (標準偏差)	最小値〜最高値	範囲	層	信頼性 Reliability
受験者	0.92 (1.44)	−6.94〜5.10	12.04	6.35	.95
評価観点	0.00 (1.54)	−3.18〜3.35	6.53	29.84	1.00
評価者	0.00 (0.17)	−0.41〜0.17	0.58	3.64	.86

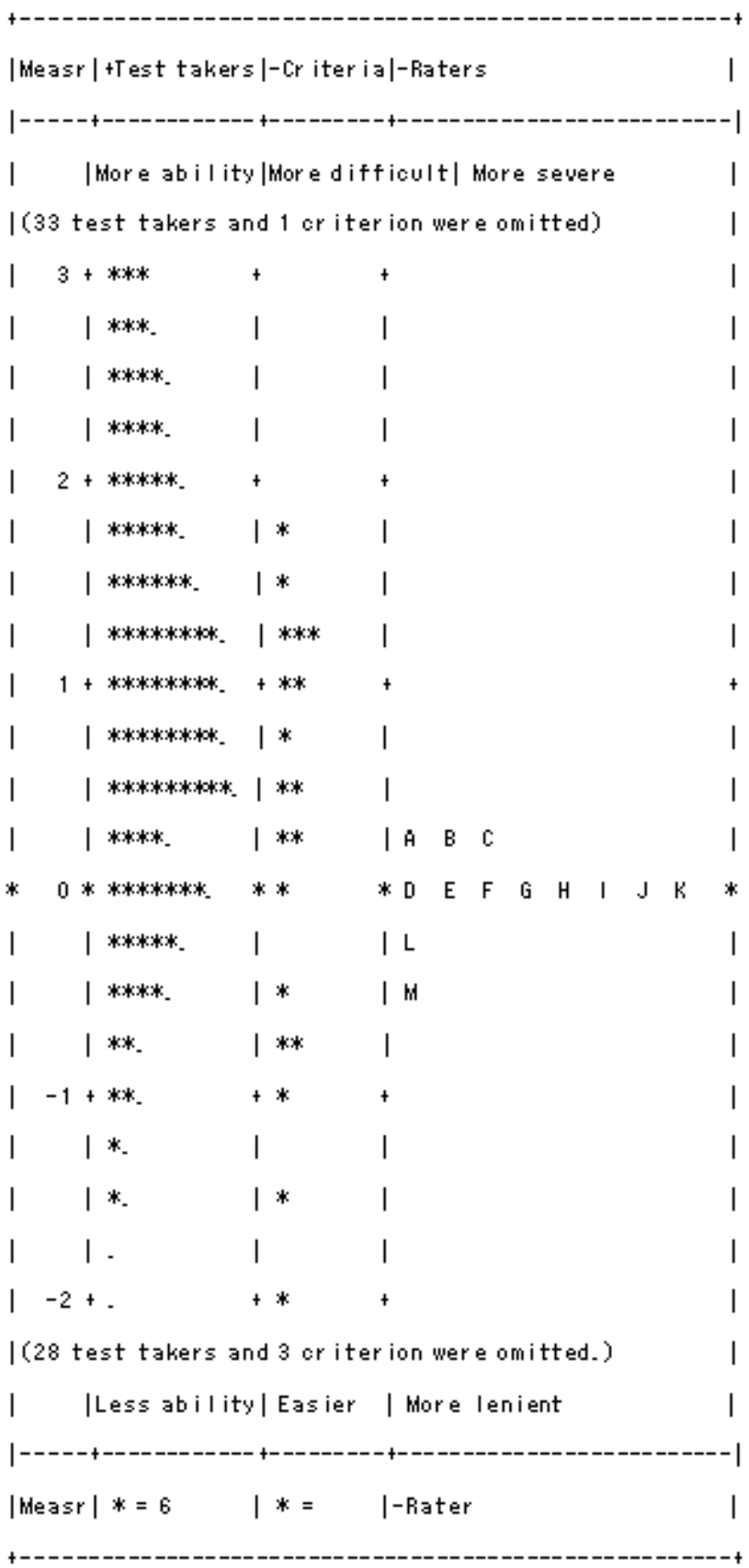

図 1. 多相ラッシュ分析の変数マップ。スペースの関係上、評価観点の category threshold 値は略。

表 2 では、上記の 2 つの基準での結果がまとめられている。どの程度モデルに適合してい

るかは使用する基準によって変わるが、受験者については、インフィット平方平均の広い基準では 8.33%と、アンダーフィットの受験者は多い傾向があった。アンダーフィットと判定された受験者は、スピーキング能力が低めだが、ある数個の難しめのタスクについては高い点を取っている場合と、スピーキング能力が高めだが、ある数個の易しいタスクについて低い点を取っている場合が考えられる。分析したところ、今回は、スピーキング能力が高めだが、ある数個の易しいタスクについて、話すべき内容を間違えて off-topic として低い点を取っている場合が多く見られた。

表 2　オーバーフィット、適合、アンダーフィットの%

		値 < 0.70 (overfit)	0.70 ≦値 ≦ 1.30 (fit)	1.30 < 値 (underfit)	値 < 0.50 (overfit)	0.50 ≦値 ≦ 1.50 (fit)	1.50 < 値 (underfit)
受験者	Infit	16.20	66.98	16.36	2.16	89.04	8.33
	Outfit	32.36	44.14	22.84	9.88	74.07	15.59
評価観点	Infit	0.00	100.00	0.00	0.00	100.00	0.00
	Outfit	4.35	73.91	21.74	0.00	100.00	0.00
評価者	Infit	0.00	100.00	0.00	0.00	100.00	0.00
	Outfit	0.00	100.00	0.00	0.00	100.00	0.00

注 : 満点を取り、フィット値が出ない受験者がおり、合計が 100%にならないところもある。

(1) 評価者の厳しさに違いはあるか

評価者の厳しさの値（Measure）を用いて検討したところ、範囲は 0.58 あり、これは、フェアスコアという、多相ラッシュ分析を用いずに点数を足し合わせた素点の尺度に直すと、0.24 点の違いがあった。受験者の範囲と比べて約 20 分の 1（0.048 =0.58/12.04）であり、厳しさの違いはあるが、小さいと考えられる。

実際の運用では、2 名の結果が不一致の場合には、第 3 の経験豊富な評価者が最終的な評価を行うが、評価者 2 名で判断する際にも、2 人とも甘い評価者や、2 人とも厳しい評価者にあたって、評価結果が一致することもあり、その場合には点数が本来の力よりも上下することはありうる。そのため、評価者訓練などをさらに向上させることが求められる。

(2) 評価者間での一致度は高いか

評価者間の一致度の%を用いて検討したところ、実際に一致したのが 79.4%だった。ラッシュ分析において一致するのが予想されるのが 63.4%であったため、それより高い結果だった。このため、評価者間信頼性は高かったと考えられる。

一致度が 79.4%であるため、採点が一致しなかったのは 20.6%の場合であった。第 1・2 の評価者の採点で、どの観点でずれていたかを調べたところ、満点が 2 点または 3 点の場合にずれが起こりがちだった。特に、2 つの理由を求めるタスクでの評価にずれが多めにみられており、評価訓練等で、より多くの評価サンプルを提示して説明や練習するなどの改善が求められる。

(3) 評価者内の一貫性は高いか

インフィット・アウトフィット平均平方で検討した。表 2 を見ると、広い基準、狭い基準の場合ともに、全ての評価者が基準内に入っており（例：インフィット平均平方の平均値 = 1.09, 標準偏差 = 0.09; 範囲 = 0.85～1.21）、評価者の行動は、ラッシュモデルから予想される範囲で一貫したものだった。そのため、評価者内信頼性は高いと考えられる。

(4) 評価者と評価観点、評価者と受験者の間に偏った評価傾向は見られるか

Facets のバイアス分析のアウトプットの *z* 値（現在のアウトプットでは *t* 値； Linacre, 2013, p. 212）を参照し、それが|±2.00|を超えたものの割合を調べた。表 3 を見ると、評価者と受験

者の組み合わせでは、2.91%で偏った評価傾向が見られた。これは、ある特定の受験者をより甘く、またはより厳しく採点していた評価者が 3%ほどいたということである。また、評価者と評価観点の組み合わせでは、25.78%で偏った評価傾向が見られた。これは、ある特定の評価観点をより甘く、またはより厳しく採点していた評価者が 26%ほどいたということである。受験者と評価観点の組み合わせでは、4.05%で偏った評価傾向が見られた。評価者と評価観点の組み合わせでの 25.78%は高く見えるが、先行研究でも、評価者訓練等の管理に力を入れたテストにおいて評価者と評価観点の組み合わせで高パーセントが報告されている。例えば、Eckes (2005) ではドイツ語テストのスピーキングセクションで 37.0%が報告されている。また、この組み合わせの偏りを減らすための方法を目的とした研究もあり（Elder, Knoch, Barkhuizen, & von Randow, 2005）、偏りが出やすく注意が必要な部分であると言える。今後この点をさらに改善するための評価者訓練やモニタリングが求められる。

表 3　バイアス分析のまとめ

	評価者 x 受験者	評価者 x 評価観点	受験者 x 評価観点
組み合わせの数	6,351	256	14,835
|±2.00|を超えた%	2.91%	25.78%	4.05%

５．おわりに

GTEC CBT の SS の評価の分析を行ったところ、全体的に好ましい結果だった。評価の厳しさには違いがあったものの、評価者間と評価者内の信頼性は共に高いものだった。評価者と評価観点の組み合わせで偏った評価結果は 4 分の 1 程度見られたが、評価者と受験者の組み合わせでは偏った評価結果では少しのみ見られた。今回の結果に基づき、評価者訓練、評価者モニタリングを改善することが求められる。また、今後は、評価者信頼性以外の観点からテストを分析し、妥当性の証拠を示していくことが必要である。

今回の分析方法は、多相ラッシュ分析を使った評価者信頼性を調べる際に、他のテストでも参考になると思われる。より詳細なテストの情報をテスト使用者に提供するために、様々なテストで妥当性検証が行われることが求められている。

６．引用文献

Akiyama, T. (2003). Assessing speaking: Issues in school-based assessment and the introduction of speaking tests into the Japanese senior high school entrance examination. *JALT Journal, 25,* 117–141.

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. III: Evaluation, Methodology, and Interdisciplinary Themes, Part 10: Quantitative analysis, pp. 1301–1322). West Sussex, UK: John Wiley & Sons. doi:10.1002/9781118411360.wbcla070

Benesse Corporation (2016). 「GTEC CBT とは　問題構成」 Retrieved from http://www.benesse-gtec.com/cbt/about/composition

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2,* 197–221.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main, Germany: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2,* 175–196.

doi:http://dx.doi.org/10.1207/s15434311laq0203_1
Hagan, S. O., Pill, J., & Zhang, Y. (2016). Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective. *Language Testing, 33,* 195–216. doi:10.1177/0265532215607920
Japan Language Testing Association (JLTA) Steering Committee & JLTA Language Testing Terminology Committee. (2006). *The JLTA bilingual list of language testing terms. The JLTA Code of Good Testing Practice.* Nagano: JLTA Secretariat. Retrieved from https://jlta.ac/?page_id=35
Koizumi, R., In'nami, Y., & Fukazawa, M. (2016). Multifaceted Rasch analysis of paired oral tasks for Japanese learners of English. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 89–106). Gateway East, Singapore: Springer Singapore. doi:10.1007/978-981-10-1687-5
Linacre, J. M. (2013). *A user's guide to FACETS: Rasch-model computer programs (Program manual 3.71.0).* Retrieved from http://www.winsteps.com/a/facets-manual.pdf
Linacre, J. M. (2014). Facets: Many-Facet Rasch-measurement (Version 3.71.4) [Computer software]. Chicago: MESA Press.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241–256. doi:10.1177/026553229601300302
Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *ARELE, 26,* 333–348.