

評価者が重視する観点とルーブリックの活用効果 —小学校外国語科における話すこと（やり取り）のパフォーマンステストにおける分析—

Rater Priorities and the Effectiveness of Rubric Use : An Analysis of Performance Tests for Speaking Interaction in Elementary School Foreign Language Classes

佐藤 剛*・大高 智英**・北向 周平**・小川 快都**・佐々木駿介**
Tsuyoshi SATO*・Tomohide OTAKA**・Shuhei KITAMUKI**・Kaito KOGAWA**・Shunsuke SASAKI**

柴田 真帆**・中村 尚平**・畠山 大輝**・柳谷 美羽**
Maho SHIBATA**・Shohei NAKAMURA**・Taiki HATAKEYAMA**・Miu YANAGIYA**

要 旨

本稿は、小学校外国語科における話すこと（やり取り）のパフォーマンステストにおいて、評価者がルーブリックを用いて評価を行う際に、重視する観点や根拠を明らかにするものである。国立大学教育学部に所属する大学生83名を対象に、佐藤他（2023）で用いられた、児童のパフォーマンスを再現したインタビューのビデオを、ルーブリックを参照しながら、総合評価および「文法の正確さ」、「発音の正確さ」、「流暢さ」、「ジェスチャー」、「アイコンタクト」、「声の大きさ」の6つの観点別に評価してもらい、重回帰分析、決定木分析によって分析した。重回帰分析の結果、佐藤他（2023）でルーブリックを用いなかった際の決定係数は.53であったのに対し、ルーブリックを用いた際の決定係数は.65であり、ルーブリックを用いると6つの要因で総合評価の6割以上を説明できることが明らかになった。また、標準化偏回帰係数の値から、「文法の正確さ」、「流暢さ」、「発音の正確さ」の順で総合評価に影響していることも示された。さらに決定木分析の結果、佐藤他（2023）では、「文法の正確さ」が優先的に見られるのに対し、ルーブリックを用いた本研究では、「流暢さ」が最も優先された。以上のことを踏まえると、評価者はルーブリックを用いない場合、発話が相手に伝わるかどうかを基準に評価を行うが、ルーブリックを用いることで、目的・場面・状況に応じた発話ができていないかどうかを基準に評価を行うと推察されるとともに、決定係数の値が上昇し評価者間の採点のぶれが減少したことから、児童のパフォーマンスを評価する際にルーブリックを用いることの重要性が示唆された。

キーワード：小学校英語教育 話すこと（やり取り） 評価

1 はじめに

近年、学校教育における英語の授業においてはコミュニケーション能力の育成の必要性が重視され、それに伴い「聞くこと」「読むこと」「話すこと（やり取り・発表）」「書くこと」の4技能5領域をバランスよく評価することが求められている。これまでのよ

うにペーパーテストで測定可能であった「読むこと」や「書くこと」とは異なり、妥当性の高い「話すこと」の評価は、学習者に何らかのタスクを与えた上で実際に話させる直接評価が不可欠である。一方で、小泉（2021）において指摘されているように、スピーキングテストは多肢選択などの客観評価での評価が難しく、主観テストでの評価が多く用いられる。評価者の

* 弘前大学教育学部

Department of English Education, Faculty of Education, Hirosaki University

** 弘前大学教育学部学校教育教員養成課程

Teacher Training Division, Faculty of Education, Hirosaki University

経験や知識、性格などによって結果が大きく異なることが多いため、信頼性の確保が大きな課題である。また、佐藤他（2023）は、評価の基準や方法を採点者に委ねた場合、文法の正確さが最も重視されること、アイコンタクトやジェスチャーなど非言語的要因に総合評価が影響されて、英語の質の評価が歪められる可能性が示唆されるなど、話すこと（やり取り）の評価には妥当性をどのように担保するかという問題を見逃できないこともまた事実であるとしている。

2 先行研究

小泉他（2017）は学習指導要領で4技能のバランスが取れた指導が求められていることから、英語の授業においてスピーキング活動が多く実施されるようになり、それに伴いスピーキングの適切な評価への需要が高まっていると示唆している。一方でスピーキングの評価は、教師の負担も大きく、作成や方法に不安を感じる教師も少なくないことをその課題としている。主な原因としては、多肢選択のような客観テストと異なり、スピーキングテストは評価者の経験や知識、性格などによって結果が大きく異なることが多いこと、言い換えれば、採点者間信頼性の問題（小泉他，2017；小泉，2021；高梨他，2023；根岸，2017）やリアルタイムで評価を行う際には、ライティングと異なり学習者のパフォーマンスの記録を残して再度評価ができない採点者内信頼性の問題が挙げられる。そのため、評価の一貫性を保つための方法として、ルーブリックを設定し評価者間で共有することが挙げられる。

金子（2022）は、コミュニケーション活動の評価において教師と生徒のやり取りの評価が多く、生徒同士のやり取りの評価の試みが少ないことに問題意識を持ち、研究を行った。勤務校で導入されており、受験者同士のやり取りを評価の対象に含めているケンブリッジ英語検定の評価方法・項目を分析し、それに対応させたスピーキング活動・スピーキングテストを授業内で行った。その結果、受験者の組み合わせが評価を左右してしまうこと、複数の評価者が採点する場合の評価者間信頼性の低さが課題として挙げられた。こうしたことから、パフォーマンステストの評価の際には、テストの目的・内容に応じたルーブリックの作成と評価者間で共通の指標を確立させて採点を行うべきであると述べている。ルーブリックを作成・共有した場合と、ルーブリック無しで評価を行った場合との評価結果の比較研究を実施することでルーブリックの作成や

共有の妥当性を示すことは、小学校外国語の授業におけるスピーキングの評価を適切に実施する上で喫緊の課題と言える。

丹藤（2023）は、アンケート調査、聞き取り調査を通して、青森県内の小学校が行ったスピーキングパフォーマンス評価の実態と課題を明らかにし、その解決策を探った。その結果、パフォーマンス評価の課題として、

- (1) パフォーマンステストにおいて、個別テストの実施率が低いこと
- (2) 評価者が1人で、その場で評価する場合が多いことから、評価の信頼性が懸念されること
- (3) 年間評価計画とルーブリックの作成率が低いこと
- (4) 採点の観点において、発音やプロソディが軽視されていること
- (5) 教師の資質・能力に関わる問題、時間や場所の物理的な問題があること
- (6) パフォーマンス評価に自信を持っている教師が多い一方で、信頼性の高い評価活動が行われているとは推察できず、そのギャップを埋める必要があること

が挙げられ、教師の資質・能力の向上を図る実践的な研修を行ったり、パフォーマンス評価の時間や場所を確保できるような労働環境の改善を図る解決策を講じたりする必要があると結論づけた。

佐藤他（2023）は小学校外国語科における話すこと（やり取り）のパフォーマンステストにおいて、ルーブリックの設定などをせず評価を採点者それぞれにゆだねた場合、採点者が重視する評価の観点や根拠を明らかにするために国立大学教育学部に所属する大学生71名を対象に、総合評価を従属変数、「文法の正確さ」、「発音の正確さ」、「流暢さ」、「ジェスチャー」、「アイコンタクト」、「声の大きさ」の6つの観点を独立変数として重回帰分析を実施した。その結果、決定係数は.53であり、標準化偏回帰係数の値から、「文法の正確さ」、「アイコンタクト」、「流暢さ」の順で評価に影響していることが明らかになった。これらの結果から小学校英語教育において、文法の正確さが評価をする上で最も必要な要因になっていること、また、非言語的要因である「アイコンタクト」が評価の大きな要因になっていることは話すこと（やり取り）の能力を測定するための評価方法としての妥当性に欠けるのではないかと問題提起している。そのため、スピーキングの評価においてはルーブリックを設定し、それを参照することが一般的である。小泉（2021）は、ルー

ブリックとは評価規準と判定基準を合わせたもので、採点をより客観的に行うためのものであるとし、その利点として以下の5つをあげている。

- (1) 採点の透明性：信頼性を高めることができ、採点のばらつきを抑えることができる
- (2) 生徒へ次の学習に向けてのフィードバックに使いやすい
- (3) 教師の指導改善の情報を得ることができる
- (4) 生徒の到達度への記録がとりやすい
- (5) 他の教師、生徒や保護者に結果や採点の意味を伝えやすい

このように、ルーブリックを設定することで、佐藤他（2023）で示された、小学校外国語教育における話すこと（やり取り）の評価における、文法の正確さへの過度な依存と、非言語的要因の影響を防ぐ可能性がある。よって、本研究は小学校外国語科における話すこと（やり取り）のパフォーマンステストにおいて、ルーブリックの有無の違いによって評価の違いを明らかにするものである。

3 リサーチクエスチョン

本稿は、小学校外国語科における話すこと（やり取り）のパフォーマンステストにおいて、ルーブリックを参照して評価する場合とルーブリック無しで評価する場合との評価結果の違いを比較するものである。そのため、以下の2つのリサーチクエスチョン（RQs）を設定した。

- RQ1 ルーブリックの有無によって小学校外国語科における話すこと（やり取り）のパフォーマンスを評価する際に評価者が重視する観点にはどのような違いがあるか。
- RQ2 ルーブリックの有無によって小学校外国語科における話すこと（やり取り）のパフォーマンスを評価する際、評価者はどのような流れで評価するか。

4 研究方法

4.1 調査協力者

本研究は、国立大学教育学部小学校コースに所属する大学2年生を対象に調査を行い、調査協力者の合計は165名である。実験参加者を、ルーブリックを参照して評価をするグループ（r+）94名とルーブリック無しで評価を行うグループ（r-）71名を分析対象とし

た。なお全員が、小学校外国語教育法の授業を履修しており、スピーキングの評価についての講義を受講済みであるため、小学校の英語のスピーキングの評価に関しての知識をもっている。

4.2 ルーブリック

本研究では、小泉他（2017）及び小泉（2021）に基づいてルーブリックの作成を行った。小泉他（2017）ではルーブリックの作成において、テストの目的が何であるのかを考えることが重要であるとしている。タスクの目的を設定せず、言語的正確さを細かく減点法で記述することは、単なる間違い探しのような評価になってしまい、より多くの英文を書いたり、新しい表現を積極的に使ったりするリスクテークした学習者の得点が低くなってしまうためである。よって、与えられたタスクに答えているかという技能面と言語的正確さである知識面を独立させて記述することを提案している。

また、ルーブリックには大きく分けて総合的（全体的）ルーブリックと分析的ルーブリックがあるが、分析的ルーブリックは採点者の負担が大きいため一般的には総合的ルーブリックが用いられることが多い（小泉，2021）。また、いずれの場合でもタスクで求められていることを達成しているかを示すタスクの達成度は含めるべきであるとしている（小泉他，2017；小泉，2021）。以上のことを踏まえて、本研究では、佐藤他（2023）と同様に小学生が「Unit 1自己紹介をしよう」の単元末に実施したパフォーマンステストを録画したビデオだという想定で評価するためのルーブリックを設定することを目的として、タスクの達成度としてALTに自分に関する情報を伝えることができる、知識面として「文法的な正確さ」、「発音の正確さ」、「流暢さ」のみを含めた記述文を作成した。佐藤他（2023）で示された課題であるアイコンタクトなど非言語的要因に英語そのものの評価が影響されることを、ルーブリックの設定によって防ぐことができるのかどうかを検証するため、アイコンタクト、ジェスチャー、声の大きさなど非言語的要因はルーブリックにあえて含めなかった。

表1 ルーブリック有のグループが評価の際に参照するルーブリック

評価	評価基準
4	文法的な間違いが全く見られず、ALT が十分に内容を理解できる自然な発音で、(自然に) なめらかに会話を続けることができる
3	文法的な間違いがほとんど見られず、ALT が十分に大まかな内容を理解できる自然な発音で、おおむね(自然に) なめらかに会話を続けることができる
2	文法的な間違いが多少見られ、ALT が内容を理解しがたいやや不自然な発音で、たどたどしさが見られる
1	文法的な間違いが多く見られ、ALT が内容を理解しがたい不自然な発音で、沈黙が目立つ

4.3 評価用動画

本研究で実験協力者が評価するパフォーマンスのビデオは佐藤他(2023)で作成したものである。このビデオは小学校での児童と教師との英語のパフォーマンステストを大学生が演じたものであり、文法の正確さ、発音の正確さ、流暢さ、ジェスチャー、アイコンタクト、声の大きさの6つの観点のいずれかが著しく劣っている状況を再現した6種類のパフォーマンスを含む。

- (1) 発音が正確かつ流暢であり、アイコンタクトを取りながらジェスチャーをつけて十分な声量で発話しているが、語順の間違いを多く含む文法の正確さが著しく劣っている児童のパフォーマンス(以下 GA)

(例) *Very much volleyball I like. Yes, do I.*

- (2) 文法的な面では正確かつ流暢であり、アイコンタクトを取りながらジェスチャーをつけて十分な声量で発話しているが、カタカナ発音での応答する発音の正確さが著しく劣っている児童のパフォーマンス(以下 PA)

(例) ハロー、アイ ライク バレーボール

- (3) 文法的・発音の面でも正確であり、アイコンタクトを取りながらジェスチャーをつけて十分な声量で発話しているが、教師から質問をされた後、答えるまでに毎回5秒程度の間が入る流暢さが著しく劣っている児童のパフォーマンス(以下 FL)
- (4) 文法的・発音の面でも正確かつ流暢であり、アイコンタクトを取りながら十分な声量で発話してい

るが、一切ジェスチャーをつけずに応答する児童のパフォーマンス(以下 GT)

- (5) 文法的・発音の面でも正確かつ流暢であり、ジェスチャーをつけて十分な声量で発話しているが、パフォーマンスを通して、一度も教師とアイコンタクトをとらない児童のパフォーマンス(以下 EY)
- (6) 文法的・発音の面でも正確かつ流暢あり、アイコンタクトを取りながらジェスチャーをつけて十分な声量で発話しているが、全体を通して声量が十分な児童のパフォーマンス(以下 VL)

本研究はルーブリックの有無によって、評価する上で重視する観点の違いを比較調査するものである。よって発話内容、発話の構成、発話の長さ、発話における統語的・語彙的複雑さなど、他の要因の影響を避けるために、話す内容は6種類すべてのビデオにおいて以下に示す共通のものとした。

教師役: Hello.

児童役: Hello.

教師役: What is your name?

児童役: I'm ○○.

教師役: How do you spell your name?

児童役: ○○○○.

教師役: What sports do you like?

児童役: I like XXX.

教師役: Can you play XXX well?

児童役: No, I can't.

教師役: Do you watch XXX on TV?

児童役: Yes, I do.

教師役: Question please.

児童役: Do you like XXX?

教師役: Yes, I do. Thank you very much.

児童役: You're welcome.

4.4 実験手続き

本実験はすべて Microsoft Teams を使ったオンデマンド形式で実施した。スピーキングの評価においては、繰り返し確認することができるように受験者のパフォーマンスを録画することが多いこと、十分な時間を確保して繰り返しビデオを見返すことができることから、オンデマンドの形式をとった。上記7種類のパフォーマンスビデオを Microsoft Teams 上で共有し、評価を課題として以下のような指示を与えた。

Teams の「ファイル」に自己紹介ビデオがアップされています。小学生が「Unit 1自己紹介をしよう」の単元末に実施したパフォーマンステストを録画したビデオだという想定で評価してください。

ルーブリックを参照して評価をするグループ ($r+$) に対し、評価規準・基準を明記したルーブリックを配布し、それを熟読した上で6種類のパフォーマンスをそれぞれ4段階で総合的に評価するよう指示をした。その後、6種類のパフォーマンスを再度見て、6つの観点別に（文法の正確さ、発音の正確さ、流暢さ、ジェスチャー、アイコンタクト、声の大きさ）4段階で評価するように指示をした。一方、ルーブリック無しで評価を行うグループ ($r-$) は、具体的な評価については評価者それぞれに委ねることを説明し、ルーブリックを参照して評価をするグループと同様に、6つのパフォーマンスをそれぞれ4段階で総合的に評価し、その後、6種類のパフォーマンスを再度見て、6つの観点別に4段階で評価してもらった。

4.5 データ分析

データ分析に当たっては、実験協力者から得たデータを csv ファイルに変換し、JASP (version 0.19.0, JASP Team, 2023) を用いて、総合評価を従属変数として、「文法の正確さ」、「発音の正確さ」、「流暢さ」、「ジェスチャー」、「アイコンタクト」、「声の大きさ」の6つの観点を独立変数として重回帰分析を行った。ルーブリックを用いずに行った実験とルーブリックを用いて行った実験のそれぞれの分析結果を比較し、ルーブリックが各観点的評価のばらつきを避けることにどのように貢献するかを調査した。また、この分析に用いた同じデータを、統計解析ソフト R (ver. 4.3.1, R Development Core Team, 2023) を用いて、決定木分析を行い、実験協力者が評価をする際の意思決定のプロセスをより詳細に分析し、その結果も比較した。

5 結果と考察

6種類のスピーキングのパフォーマンスに対する4段階の評価の平均値、標準偏差、最小値、最大値は以下の表2のとおりである。6種類のパフォーマンスについては、ルーブリックを参照したグループ ($r+$)、ルーブリック無しで評価するグループ ($r-$) とともに2.97から3.61を平均に評価者によって大きなばらつきがあり、4の評価をした評価者から1の評価をつけた評価者まで大きな幅があることが分かる。

表2 記述統計量

		M	SD	Min	Max
GA	$r-$	3.37	0.99	1.00	4.00
	$r+$	3.49	1.03	1.00	4.00
PA	$r-$	3.22	0.92	1.00	4.00
	$r+$	3.35	0.82	1.00	4.00
FL	$r-$	3.03	1.03	1.00	4.00
	$r+$	3.11	1.01	1.00	4.00
GT	$r-$	3.07	1.07	1.00	4.00
	$r+$	2.97	1.12	1.00	4.00
EY	$r-$	3.22	1.07	1.00	4.00
	$r+$	3.37	1.05	1.00	4.00
VL	$r-$	3.47	0.68	1.00	4.00
	$r+$	3.61	0.85	1.00	4.00

注： $r+$ =ルーブリック有、 $r-$ =ルーブリック無、 M =平均値、 SD =標準偏差、 Min =最小値、 Max =最大値

5.1 重回帰分析によるルーブリックの有無による評価の違い

総合評価に対する6つの観点（文法の正確さ、発音の正確さ、流暢さ、ジェスチャー、アイコンタクト、声の大きさ）の影響を検討するために、ルーブリック無しで評価するグループとルーブリックを参照して評価するグループそれぞれ重回帰分析を行った。具体的には総合評価を従属変数、6つの観点を独立変数として分析を行った。以下の表3はルーブリックなしで評価するグループの結果を、表4はルーブリックを参照して評価するグループの結果をそれぞれ示している。

表3 総合評価を従属変数にした重回帰分析の結果（ルーブリック無）

変数	B	SEB	β
文法の正確さ	0.34	0.02	0.43**
発音の正確さ	0.14	0.03	0.16**
流暢さ	0.17	0.03	0.23**
ジェスチャー	0.13	0.02	1.18**
アイコンタクト	0.28	0.02	0.38**
声の大きさ	0.13	0.03	0.15**
切片	-1.07	0.02	

注： $R^2 = .53$, ** $p < .001$, B = 偏回帰係数, SEB = 標準誤差, β = 標準化偏回帰係数

ループリック無しで評価を行うグループでは独立変数である6つの観点（文法の正確さ、発音の正確さ、流暢さ、ジェスチャー、アイコンタクト、声の大きさ）はいずれも有意であり、その標準化偏回帰係数は、文法の正確さ（ $\beta = 0.43$ ）、アイコンタクト（ $\beta = 0.38$ ）、流暢さ（ $\beta = 0.23$ ）、ジェスチャー（ $\beta = 0.18$ ）、発音の正確さ（ $\beta = 0.16$ ）、声の大きさ（ $\beta = 0.15$ ）の順に高い結果となった。このことから、実験協力者は主に文法的な正確さとアイコンタクト、流暢さを基準にスピーキングの評価をしていることが示された。

決定係数 $R^2 = .53$ であり、分散分析による有意性検定の結果、有意であることが示された（ $F(6, 601) = 114.34, p < .001$ ）。このことから、上記6つの独立変数により、従属変数である総合評価の53%を説明できることが明らかになった。

表4 総合評価を従属変数にした重回帰分析の結果（ループリック有）

変数	B	SEB	β
文法の正確さ	0.51	0.03	0.55**
発音の正確さ	0.19	0.04	0.17**
流暢さ	0.38	0.03	0.41**
ジェスチャー	-0.03	0.03	-0.04
アイコンタクト	0.10	0.03	0.11**
声の大きさ	0.03	0.03	0.02
切片	-0.94	0.24	

注： $R^2 = .65$, ** $p < .001$, B = 偏回帰係数, SEB = 標準誤差, β = 標準化偏回帰係数

ループリックを参照して評価をする場合、6つの観点のうち有意な独立変数は文法の正確さ、発音の正確さ、流暢さ、アイコンタクトの4つであり、その標準化偏回帰係数は、文法の正確さ（ $\beta = 0.55$ ）、流暢さ（ $\beta = 0.41$ ）、発音の正確さ（ $\beta = 0.11$ ）、アイコンタクト（ $\beta = 0.11$ ）の順に高い結果となった。一方でジェスチャーと声の大きさの標準化回帰係数はそれぞれジェスチャー（ $\beta = -0.04$ ）、声の大きさ（ $\beta = 0.11$ ）であり、有意ではなかった。このことから、ループリックを参照して評価する場合、実験協力者はジェスチャーや声の大きさなど非言語的要因に左右されず、文法の正確さ、流暢さ、発音の正確さなど言語的要因に基づいて評価を行う傾向が高くなることが示された。また、決定係数 $R^2 = .65$ であり、分散分析による

有意性検定の結果、有意であることが示され（ $F(6, 510) = 318.44, p < .001$ ）、上記6つの独立変数により、従属変数である総合評価の65%を説明できることが明らかになり、ループリック無しで評価する場合と比較して、6つの独立変数によって説明できる従属変数の割合が高くなったことが分かる。

上記4.2に示す通り、本研究で作成したループリックの記述文に含めたものは「文法的な正確さ」、「発音の正確さ」、「流暢さ」など言語面に関する要因であり、非言語的要因に英語そのものの評価が影響されることを、ループリックの設定によって防ぐことができるのかどうかを検証するため、アイコンタクト、ジェスチャー、声の大きさなど非言語的要因はあえて含めなかった。その結果として、非言語的要因であるジェスチャーや声の大きさが有意な独立変数とならなかったことは、ループリックが評価の妥当性を高める可能性を示すひとつの結果と言える。ただ、一方で非言語的要因のひとつであるアイコンタクトが、標準化偏回帰係数は0.38から0.11に低くなったものの、従属変数に影響する有意な要因となっている点は、特筆すべき点である。採点者はループリックに記述されていないにもかかわらず、アイコンタクトの有無によって評価を変えているということである。この原因として、1998年改訂の学習指導要領で総合的な学習の一環として実施されてきた外国語に触れたり、外国の生活・文化に慣れ親しんだりする体験的学習以来、これまで小学校においてコミュニケーションを非言語的要因を広く含めて指導されてきたことが挙げられると考えられる。特にアイコンタクトにとっては、めあてのひとつに取り上げるなどそのほかの要因と比較して重要されてきた指導項目であることが影響している可能性がある。授業のめあてとして、普段の授業で指導されてきたものであり、ループリックに評価規準のひとつとして記述されているのであれば問題はないが、本研究で作成したループリックのようにアイコンタクトに該当する記述がない場合に、アイコンタクトの有無が総合評価に影響を与えることは避けるべきである。

5.2 決定木分析によるループリックの有無による評価の違い

決定木分析は、代表的な「教師あり学習」の機械学習の手法であり複数の説明変数をもとにして、情報を得ようとする目的で使用される分析方法であり、求め

られる結果がそれぞれの変数によって分岐されたものであるため、解釈が容易である利点がある（内田他, 2023）。上記の重回帰分析と同様に、総合評価を従属変数、6つの観点（文法の正確さ、発音の正確さ、流暢さ、ジェスチャー、アイコンタクト、声の大きさ）を独立変数として、ループリックを参照して評価するグループと、ループリック無しで評価するグループそれぞれに決定木分析を行った。図1はループリック無しで評価するグループの結果を、図2はループリックを参照して評価するグループの結果を示している。

図1から決定木の解析により、「文法の正確さ」によって最初の分岐が形成されていることが分かる。この分岐により、「文法の正確さ」が3.5未満の下位グループと3.5以上の上位グループに分類されている。さらに両グループとも「アイコンタクト」の変数で再分岐していることが分かる。分岐の基準は「文法の正確さ」が3.5未満の下位グループは2.5であり、3.5以上の上位グループは3.5である。さらに、下位のグループは、再度「文法の正確さ」で分岐し、最終的に「アイコンタクト」と「流暢さ」の変数で分岐していることが分かる。一方、上位グループにおいて「発音の正確さ」、「ジェスチャー」、「声の大きさ」、「流暢さ」を基に、さらに細かいサブグループが形成されている。ループリック無しで評価をする場合、評価者は、「文法の正確さ」を基準に上位と下位を分けた後、「アイコンタクト」および「文法の正確さ」を基準に評価を繰り返していることが分かる。さらに、それ以外の「発音の正確さ」、「ジェスチャー」、「声の大きさ」の

観点は上位グループを評価する際のみ用いられ、下位グループの評価には「文法の正確さ」、「アイコンタクト」、「流暢さ」のみをもとに評価している点も注目すべき点である。言い換えると、ループリック無しで評価をする場合、評価者は、まず「文法の正確さ」で上位と下位を分けて、その後、下位のグループを「アイコンタクト」、「文法の正確さ」、「流暢さ」で評価する。上位グループは「発音の正確さ」、「ジェスチャー」、「声の大きさ」に基づいてさらに細かく評価を変えていることが分かる。

本研究は小学校外国語の評価に関するものであることを考慮すると、「文法の正確さ」がまず1つ目の分岐となってしまっていること、下位グループを評価する際に再度分岐の要因となるなど、「文法の正確さ」が評価の際の重要な要因となっていることは、大きな問題といえる。この結果は外国語を評価する際に評価者は文法の正確さに自身が思う以上にとらわれる傾向があることを認識する必要性を示唆するものであり、必要以上に重視しないことが重要である。

さらに、非言語的要因である「アイコンタクト」の観点が評価において重要視されている点にも注意する必要がある。相手にメッセージを伝える上でアイコンタクトの重要性を否定するものではないが、小学校外国語の話すこと（やり取り）の評価において、アイコンタクトが言語的要因である「発音の正確さ」や「流暢さ」よりも優先されてしまうことについての妥当性は慎重に検討されるべきことであろう。

それでは、評価者がループリックを参照して話すこ

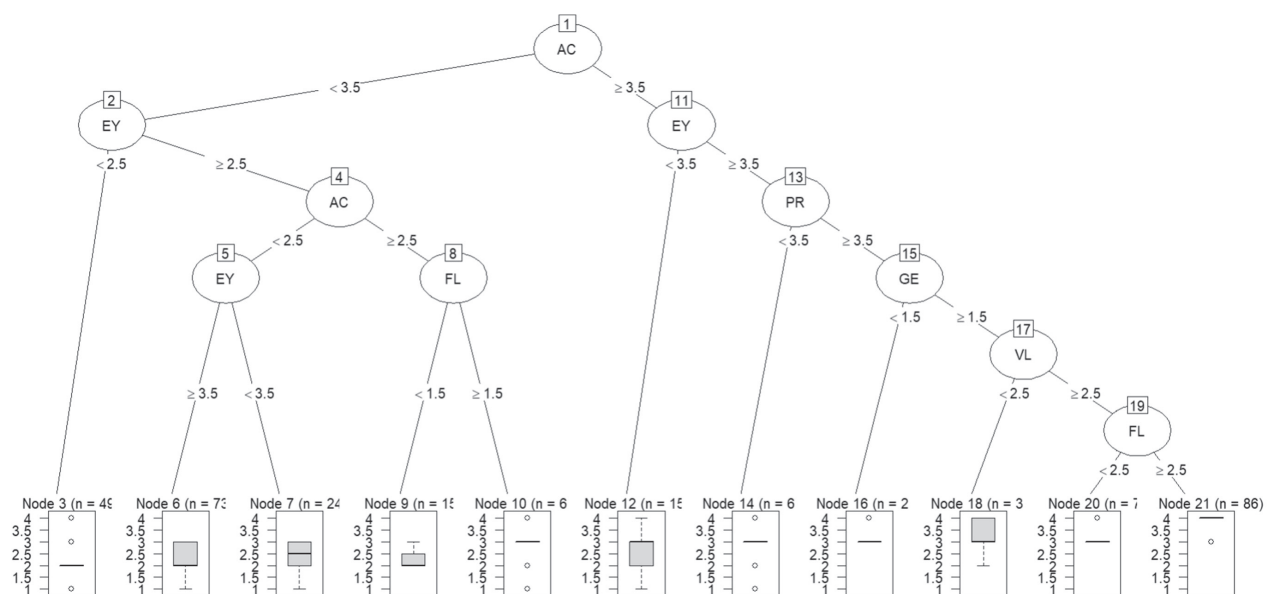


図1 総合評価を従属変数にした決定木分析の結果（ループリック無）

と（やり取り）の評価をした場合、どのように評価するのであろうか。決定木分析の結果は以下の図2に示すとおりである。

分析の結果、決定木は「流暢さ」の観点によって最初に分岐し、3.5を基に上位グループと下位グループの2つの主要な分岐が形成されている。上位グループでは、次に「文法の正確さ」の観点に基づく分岐があり、閾値1.5を満たす場合に最も高い評価である4と評価される傾向が見られた。一方、下位グループ内では、「流暢さ」の観点と「ジェスチャー」の観点がさらなる重要な分岐点として識別され、それらの観点を基にサブグループが形成されている。

ループリックを参照して評価する場合は、ループリック無しで評価する場合と比較して、決定木全体がシンプルであることが分かる。ループリックにより、着目すべき観点が明確化され、どのようなパフォーマンスが望ましいのかについてのイメージを持った状態で評価することができていると推測される。また、決定木の分岐は「流暢さ」、「文法の正確さ」、「ジェスチャー」の観点であり、ループリック無しで評価するグループに見られた、「アイコンタクト」、「声の大きさ」、「発音の正確さ」の観点が分岐の要因として表れていないことも特筆すべき点である。ループリックがあることで、これらの要因の影響を受けずに評価することにつながっていることが示唆されている。

さらに「ジェスチャー」の観点は1.5を閾値に、数値が高い場合評価が下がることも重要な結果である。つまり「ジェスチャー」の観定の点数が高いほど、総合評価が低いことを示している。意思の伝達に必要で

はない不自然に大きなジェスチャーはむしろマイナスに影響するということであり、評価者が非言語的要因よりも言語的要因に基づいて評価していることを示している結果であると考えられる。

最後に、ループリックに観点としてあるにも関わらず、「発音の正確さ」が分岐の要因として出現しなかったことが挙げられる。その原因として、小学校段階ではあまり発音に対して過度な修正をしない傾向にあること、現在の発音の指導が World Englishes の考え方に基いて、ネイティブスピーカーの英語だけが絶対的に正しいとするのではなく、母語のアクセントの影響を受けた英語の多様性を容認しようという流れにあること、そして最後に今回設定した ALT が内容を理解できるかどうかという重点に置いたことが挙げられる。上述の通り、今回、作成したパフォーマンス動画において、発音が著しく劣っているパフォーマンスは、ハロー、アイ ライク バレーボールのように極端なカタカナ読みのものであった。一言に ALT といっても日本語学習歴や日本の滞在歴などによって差はあるものの日本で生活していることは事実であるためカタカナ英語は許容されるものと評価者が判断した可能性が高い。ALT が理解できるかどうかという記述では、「発音の正確さ」という観定に評価者の注意が向かないことが示唆された結果になった。

6 まとめと教育的示唆

本稿は、小学校外国語科における話すこと（やり取り）のパフォーマンステストにおいて、評価者がルー

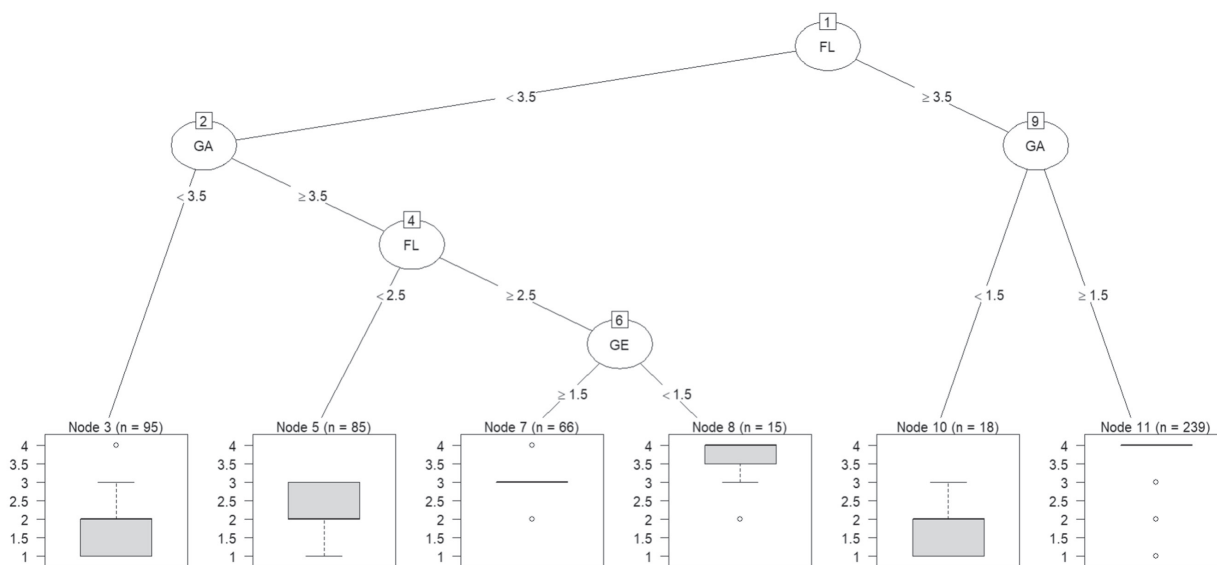


図2 総合評価を従属変数にした決定木分析の結果（ループリック有）

ブリックを参照して評価を行う場合とルーブリックを参照せずに評価する場合に重視する観点や根拠などの違いを明らかにするものである。総合評価を従属変数に「文法の正確さ」、「発音の正確さ」、「流暢さ」、「ジェスチャー」、「アイコンタクト」、「声の大きさ」の6つの観点別評価を独立変数として重回帰分析、決定木分析によって分析した。重回帰分析の結果、ルーブリックを用いなかった際の決定係数は.53であったのに対し、ルーブリックを用いた際の決定係数は.65であり、ルーブリックを参照することによって6つの要因で総合評価の説明できる割合が高くなることが示された。また、標準化偏回帰係数の値から、「文法の正確さ」、「流暢さ」、「発音の正確さ」の順で総合評価に影響していることも示された。さらに決定木分析の結果、ルーブリック無しで評価する場合には、「文法の正確さ」が最も総合評価に影響を与える要因であったのに対し、ルーブリックを参照した場合では、「流暢さ」が最も強い要因である結果となった。以上のことから、ルーブリックを用いることで、評価すべき点としなくてもよい点の差別化がなされ、評価の妥当性と信頼性の向上が期待できることが示され、小学校外国語教育において話すこと（やり取り）の評価をする際のルーブリックの有効性を実証する結果となった。

本研究から得られた教育的示唆としては、ルーブリックを作成しそれを参照することで、評価基準に含まれない要因に左右されずに、評価基準にフォーカスした評価がある程度可能になることが示されたこと、非言語的要因の総合評価に対する影響を抑制することで英語の質による評価をすることにつながり、英語の話すこと（やり取り）の評価として妥当性を高めることができること、一方で、評価者は文法の正確さやアイコンタクトに影響される危険性があることが挙げられる。

まとめとして話すこと（やり取り）の的確な評価のためのルーブリックの作成と活用においては、評価者がそのルーブリックにおいて求められている評価基準を明確に示す記述文が重要である。さらに日本人の英

語教師と外国人英語指導助手など複数での評価者間で事前にルーブリックに関する十分な話し合いを行いコンセンサスを得ることも重要であるが、事後に実際にルーブリックで示された通りの観点に基づいた評価がなされているのか検証をすることが不可欠であることが示された。本研究で取り入れた重回帰分析や決定木分析などは事後の検証においてルーブリックの見直しと改善のための非常に有益な情報を提供するものである。本研究で実施した研究手法そのものが、結果と併せて小学校外国語教育における話すこと（やり取り）の評価の改善の一助となることを期待するものである。

参考文献

- 内田治・佐野夏樹・佐野雅隆・下野僚子（2023）.『実習R言語による多変量解析：基礎から機械学習まで』サイエンス社
- 金子麻子（2022）.「ケンブリッジ英語検定の導入とそのスピーキング活動への応用－新学習指導要領「話すこと[やりとり]」の指導と評価の試み－」『研究紀要』第67号, 81-100.
- 小泉利恵・印南洋・深澤真（2017）.『実例でわかる英語テスト作成ガイド』大修館書店
- 小泉利恵（2021）.『実例でわかる英語スピーキングテスト作成ガイド』大修館書店
- 佐藤剛・内海里奈・大島梨理香・大高智英・北向周平・佐々木駿介・中村尚平・畠山大輝（2023）.「小学校外国語科の話すこと（やり取り）の評価を左右する要因は何か」『弘前大学教育学部紀要』第130号, 51-58
- 高梨庸雄・高橋正夫・佐藤剛・野呂徳治・粕谷恭子・田縁真弓（2023）.『新・英語教育学概論』金星堂
- 丹藤永也（2023）.「青森県内小学校におけるスピーキングパフォーマンス評価の実態と課題」『東北英語教育学会研究紀要』第43号, 84-96
- 根岸雅史（2017）.『テストが導く英語教育改革：「無責任なテスト」への処方箋』三省堂
- JASP Team (2024). JASP (Version 0.19.0)[Computer software].
- R Development Core Team. (2023) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

(2024. 8. 28 受理)