

話すこと【やり取り】におけるパフォーマンス評価 —教師の主観的評価と CAF による客観データの比較—

Performance Evaluation in Speaking and Interaction: Comparing Subjective Teacher Assessments with Objective CAF Framework Analysis

佐 藤 剛*・野 呂 香**
Tsuyoshi SATO, Kaori NORO

要旨

この実践報告は、68名の中学生を対象に英語コミュニケーションのパフォーマンステストを行い、教師による評価とCAF（Complexity, Accuracy, Fluency）に基づく分析データを組み合わせる評価の試みを検証するものである。結果、評価基準として「流暢さ」が最も重要であり、一方で「語彙的複雑さ」が高いほど評価が低くなる傾向が示された。CAFによる分析は、教師の評価と概ね一致しており、客観的データが教師の評価支援に有用であることが確認された。しかし、分析にかかる労力やCAF指標が生徒にとって理解しづらい点が課題として示された。特に、基準を設定し、評価結果を生徒が理解しやすい形で提供することが求められる。最終的に、CAFの活用は年間数回に限定し、教師自身の評価を見直すための手段として活用することが現実的であるという結論が得られた。

キーワード：評価、話すこと（やり取り）、CAF

1. 実践の背景

現行の学習指導要領（文部科学省、2017）では、4技能5領域（「聞くこと」、「話すこと（発表）」、「話すこと（やり取り）」、「読むこと」、「書くこと」）をバランスよく指導し、評価することが求められている。この点に関して、本学附属小中学校の教員との協同研究定例会で共通の課題として挙げられたのが、「話すこと」に関する評価の困難さである。具体的には、成績をつける際、どのような評価方法で、どのような基準に基づいて評価したのかを児童生徒はもちろん、保護者にも必要に応じて説明する責任が求められる。しかし、他の技能と比較して「話すこと」の評価には以下のような困難が伴うと指摘された。

まず、「話すこと」の評価は、ペーパーテストを用いて評価する他の技能と比較して歴史が浅く、評価方法が十分に確立されていない。また、それによって評価者が不安を感じることも課題として挙げられた。次に、ペーパーテストと異なり、生徒の解答が形として残らないことも課題である。録音や録画が可能な状況であれば別だが、解答用紙に書かれた答えを基に採点基準を明示できる他の技能の評価と比較すると、成績差をどのように示すのかが課題として挙げられた。さらに、「話すこと」の評価は多肢選択形式や正解を特定するような客観的テスト形式を採用することが難しく、主観評価が主となる。そのため、ルーブリックの活用などの工夫を行ったとしても、その客観性やアカウンタビリティを担保することが非常に困難であるとの意見が出された。

このような背景から、「話すこと」の評価において、教師による主観評価に加え、スピーキング研究分野で採用されている客観的なデータ分析手法を組み合わせる評価方法が着想された。具体的には、生徒のパフォーマンスをビデオ録画し、その文字起こしデータを基にスピーキングの熟達度を測定するため、複雑さ、正確さ、流暢さ（CAF）の概念に基づいて数量的データを算出する。このCAFの概念はスピーキング能力

* 弘前大学教育学部英語教育講座 Department of English, Faculty of Education, Hirosaki University

** 弘前大学教育学部附属中学校 Hirosaki University Faculty of Education Junior High School

の分析で多くの研究に導入されており、発話能力とCAF各要素が強く相関していることが示されている (Najmi et al., 2021; Foster & Skehan, 1996; Housen & Kuiken, 2009)。複雑さは、言語の構文や語彙の精緻さ、豊かさ、多様性、変化を反映する (Foster & Skehan, 1996; Housen & Kuiken, 2009)。正確さは「特定の基準からの逸脱の度合い」や「誤りのない状態」に焦点を当てる (Housen & Kuiken, 2009, p.463; Foster & Skehan, 1996, p.304)。流暢さは、「話すことや書くことの容易さ、雄弁さ、スムーズさ」に関連する (Housen & Kuiken, 2009, p.463)。

本稿では、教師の主観的評価にCAFを用いた客観データを組み合わせる実践を報告し、その成果と課題を検討する。これにより、より妥当性と信頼性の高い「話すこと（やり取り）」のパフォーマンス評価の可能性を模索し、今後の方向性を探ることを目的としている。

2. 実践の概要

2.1 参加者

本実践報告の参加者は、国立大学教育学部附属中学校の2クラスに所属する生徒68名である。参加者は小学校低学年から外国語活動の授業を受けており、公立中学校の生徒と比較して英語熟達度が高い。また、附属中学校に通う生徒であるため、日常的に英語に触れる機会が多く、英語を話すことに対する抵抗が少ない。さらに、コミュニケーション活動に積極的に取り組む生徒が多い点が特徴である。本研究には、平均的な中学生と比較して例外的に高い英語力を持つ生徒も含まれているが、通常授業内で実施された実践であるため、これらの生徒を分析対象から除外することはしなかった。

2.2 調査方法

分析対象とするパフォーマンステストは、「学級担任に何をプレゼントするか」についてペアで話し合い、合意形成を図るタスクで実施された。生徒のパフォーマンス評価には、以下の図1に示すルーブリックを用い、中学校教員と大学教員が観点別に分担して評価を行った。その後、これらの評価を合算し、中学校教員が最終評価を行う形を採用した。

※活動を振り返って、二重囲みの部分を自己評価してみましょう。(赤ペンを使って○をつけましょう。)

| | 観点 | 評価規準 | 評 価 | | |
|----------|------------------|-----------------------------------------|-----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|--------------|
| | | | A | B | C |
| 1 思考判断表現 | 自分の考えと理由を伝えたか。 | トピックに対する自分の考えと明確な理由を伝えている。 | トピックに対する 自分の考えと明確な理由 を伝えている。 | トピックに対する自分の考えは伝えているが、理由があいまいである。 (because it is good / because I like it など) | B基準を満たしていない。 |
| 2 思考判断表現 | やりとりが効果的か。 | 相手の発言を踏まえて話し合っている。 | 相手の発言を踏まえた上で、自分の主張をさらに述べ合意形成を図ろうとしている。 | 一方的に合意形成を図ろうとしており、効果的な話し合いをしていない。 | B基準を満たしていない。 |
| 3 思考判断表現 | タスクを達成したか。 | 話し合いを通して、合意形成をしている。 | 1つの案にまとめ、 これからの行動につながる具体的な話 をしている。 | 1つの案にまとめ、合意形成ができている。 | B基準を満たしていない。 |
| 4 主体的態度 | ポートフォリオを活用しているか。 | ポートフォリオを記入することで自分の活動を振り返り、今後に生かそうとしている。 | 自分の活動を振り返って、 具体例 を入れながらポートフォリオに記入し、 今回の活動とのつながり を意識して、次回に生かそうとしている。 | 自分の活動を振り返って、ポートフォリオを記入し、次回に生かそうとしている。 | B基準を満たしていない。 |
| 5 知識技能 | 正確さ 流暢さ | おおむね正確で、不自然な間を置かず話している。 | 意味の伝達に関係する間違いがほぼなく、不自然な間を置かず話している。 ・データ分析 正確さ目安 75%以上 流暢さ目安 WPM75～200 未満 ・観察による評価 | 間違いがあるものの、コミュニケーションが成立し、途切れながらもやり取りが続いている。 ・データ分析 正確さ目安 50%以上 流暢さ目安 WPM50～75 未満 or 200 以上 ・観察による評価 | B基準を満たしていない。 |

注. 実線で囲まれた部分：中学校の教員がパフォーマンスを見ながらその場で評価する項目、点線で囲まれた部分：大学がCAFのデータに基づいてデータを定評する項目

図1 パフォーマンステスト評価のためのルーブリック

中学校の教員は、主に図1の実線で囲まれた「思考力・判断力・表現力」と「主体的に学習に取り組む態度」の観点を、以下の基準に基づいて評価を行った。

1. トピックに対する自分の考えと明確な理由を伝えているか。
2. 相手の発言を踏まえて話し合っているか。
3. 話し合いを通して、合意形成をしているか。
4. ポートフォリオを記入することで自分の活動を振り返り、今後に生かそうとしているか。

本実践では、大学と協力して評価を行うことで教師の負担軽減を図ることも目的の一つである。そのため、教員は主にその場で評価を行い、ビデオの参照は必要最低限（再確認が必要な場合など）に留めるようにした。

一方、大学では、生徒のパフォーマンスビデオを参照しながら、CAF（Complexity, Accuracy, Fluency）に基づく数値を算出するため、以下の作業を実施した。

1. 生徒一人一人の発話時間を計測
2. 生徒の発話内容の書き起こし
3. 書き起こしデータをASユニットに分割
ASユニット（Analysis of Speech Unit）は、Foster et al.（2000）に基づく分析単位で、不完全発話やイントネーション、間、対話における中断や支援を含む話し言葉特有の特徴を反映する
4. 発話データのエラー箇所を特定
5. 総語数および異なり語数（Type-token ratio）の算出

その後、図1の点線で囲まれた「知識・技能」の観点に基づき、以下の4つの評価データを中学校教員へ提供した。

正確さ：

「エラーのないASユニットの割合」として評価

例：

発話：*I like soccer. It is exciting. I want to be soccer player in future. I like to practicing the soccer very well.*

ASユニット数：4

エラーのないASユニット数：2（例：soccer playerにおける冠詞の欠落、to practicingの誤り）

正確さ： $2 \div 4 \times 100 = 50\%$

統語的複雑さ：

ASユニット内の節数の割合で評価

例：

発話（1）*I have a friend and he can play soccer well.*

ASユニット数：2

節数：2

複雑さ： $2 \div 2 = 1$

発話（2）*I have a friend who can play soccer well.*

ASユニット数：1

節数：2

複雑さ： $2 \div 1 = 2$ （（1）の2倍の複雑さ）

語彙的複雑さ：

総語数に占める異なり語数 (Type-token ratio) として評価

例：

発話 (3) *I like soccer. I like tennis. I like baseball.*

総語数：9

異なり語数：5

語彙的複雑さ： $5 \div 9 \times 100 = 55\%$

発話 (4) *I like soccer. It's exciting. I am a good player.*

総語数：10

異なり語数：9

語彙的複雑さ： $9 \div 10 \times 100 = 90\%$

流暢さ：

1 分間に発話した単語数 (Word Per Minute, WPM) として評価

算出方法：児童生徒の発話語数 ÷ 発話時間

3. 結果と考察

生徒のスピーキングのパフォーマンスに対する評価の平均値，標準偏差，最小値，最大値，95%信頼区間は以下の表1の通りである。いずれの項目においても，平均値が高く，参加生徒は非常に高いレベルのパフォーマンスを示していることが分かる。(統語的複雑さ：136.82，語彙的複雑さ：70.30，正確さ：70.22，流暢さ：75.73，教師による最終評価：2.46)。以下の分析を考察する際には，この点を考慮する必要がある。

表1 評価データの記述統計量 (N = 62)

| | S-complexity | L-complexity | Accuracy | Fluency | Final Grade |
|---------|----------------|---------------|---------------|---------------|-------------|
| Mean | 136.82 | 70.30 | 70.22 | 75.73 | 2.46 |
| SD | 20.61 | 11.79 | 15.14 | 30.51 | 0.56 |
| Minimum | 100.00 | 45.00 | 40.00 | 15.91 | 1.00 |
| Maximum | 200.00 | 100.00 | 100.00 | 186.00 | 3.00 |
| 95% CI | [131.69-141.9] | [67.37-73.24] | [66.46-73.99] | [68.14-83.33] | [2.32-2.60] |

注. SD = 標準偏差，S-complexity = 統語的複雑さ，L-complexity = 語彙的複雑さ，Accuracy = 正確さ，Fluency = 流暢さ，Final Grade = 教師による最終評価

教師による最終評価に対する4つの客観的指標（統語的複雑さ，語彙的複雑さ，正確さ，流暢さ）の影響を検討するため，重回帰分析を実施した。具体的には，教師による最終評価を従属変数とし，4つの客観的指標を独立変数として，決定係数 R^2 および標準化回帰係数 β の検討を行った。その結果は以下の表2に示すとおりである。

独立変数である客観的指標（統語的複雑さ，語彙的複雑さ，正確さ，流暢さ）の標準化回帰係数はそれぞれ，統語的複雑さ ($\beta = 0.05$)，語彙的複雑さ ($\beta = -0.04$)，正確さ ($\beta = 0.33^*$)，流暢さ ($\beta = 0.50^{**}$)であり，有意な独立変数は正確さおよび流暢さの2つであった。このことから，教師による最終評価は主に正確さと流暢さに基づいて行われていることが示された。つまり，教師は流暢かつ正確なやり取りを行う生徒を高く評価している一方で，統語的および語彙的複雑さの標準化回帰係数は低く，どれだけ複雑な英語を使

用しても教師による最終評価には影響を与えないことが分かる。

決定係数 $R^2 = .37$ 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 ($F(4, 61)$, $p < .001$)。このことから、上記4つの独立変数によって、従属変数である教師による最終評価の37%を説明できることが明らかとなった。言い換えれば、残りの63%は他の要因によって最終評価が決定されていると考えられる。

表2 教師による最終評価を従属変数にした重回帰分析の結果

| | B | SE | β | t | p |
|--------------|------|------|---------|-------|--------|
| (Intercept) | 0.87 | 0.68 | | 1.27 | .209 |
| S-complexity | 0.00 | 0.00 | 0.05 | 0.43 | .664 |
| L-complexity | 0.00 | 0.00 | -0.04 | -0.36 | .717 |
| Accuracy | 0.01 | 0.00 | 0.33 | 3.1 | .003 |
| Fluency | 0.01 | 0.00 | 0.50 | 4.46 | < .001 |

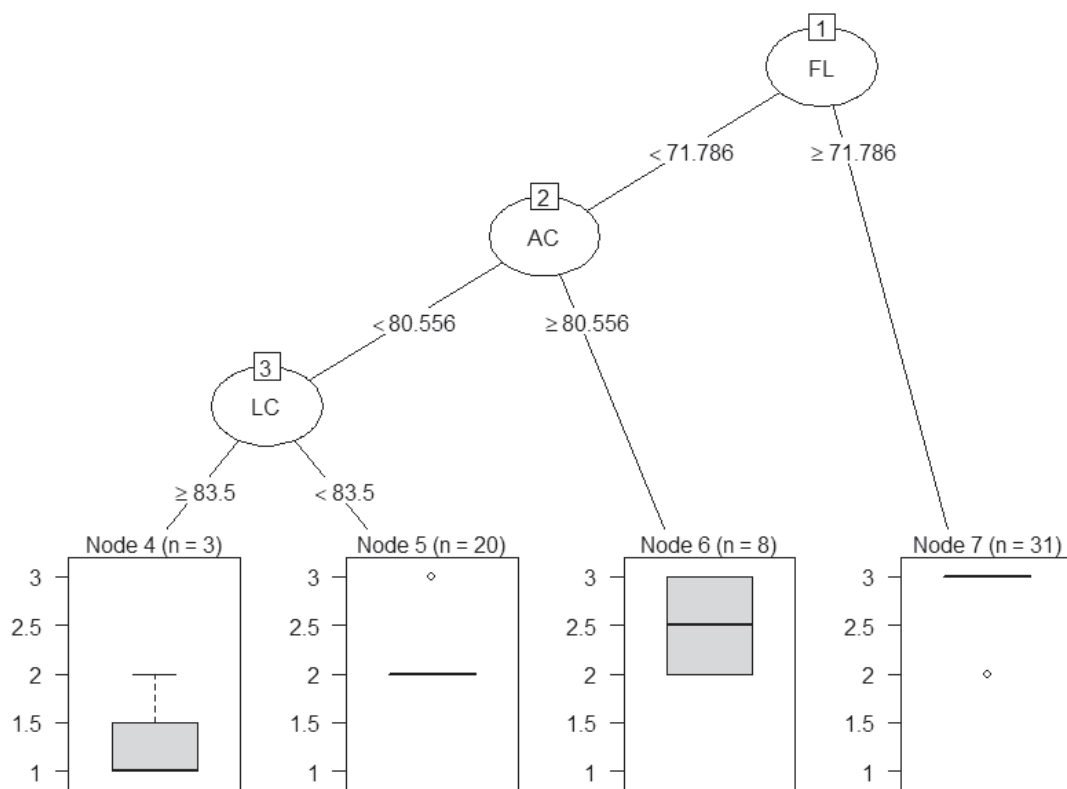
注. $R^2 = .37$, B = 偏回帰係数, SE = 標準誤差, β = 標準化回帰係数

本実践の目的は、生徒の話すこと【やり取り】の評価をする際に、教師の主観的評価とCAFによる客観データの比較および統合であるので、4つの客観的指標（統語的複雑さ、語彙的複雑さ、正確さ、流暢さ）で教師による最終評価のすべてを説明することは、必要ないと考えられる。実際に指導と評価を担当した教師にこの結果を伝え、意見を求めた結果以下のようなフィードバックを得た。

普段は意識していなかったが、データとして示されると、今回のタスクは生徒同士の会話であるため、指導においても評価においてもスムーズかつ相手に伝わることを重視した。そのため、流暢さとある程度の正確さに基づいて評価を行っているという点には納得感がある。また、同様に、生徒同士による会話の際には、必要以上に複雑な言い回しや表現などは避け、お互いが理解できる英語を使うように普段から指導している。このことが、統語的複雑さおよび語彙的複雑さが有意な独立変数とならなかった原因であると考えられる。

決定係数が37%ということは、統語的複雑さ、語彙的複雑さ、正確さ、流暢さ以外にも評価の際に考慮している要因があることを示唆しており、例えば「日ごろの生徒の様子」や「テストに向かって努力している様子」などがその要因として挙げられると考えられる。

決定木分析は、代表的な「教師あり学習」の機械学習手法であり、複数の説明変数を基に情報を得る目的で使用される分析方法である。求められる結果がそれぞれの変数によって分岐されるため、解釈が容易であるという利点がある（内田他, 2023）。上記の重回帰分析と同様に、教師による最終評価を従属変数、4つの客観的指標（統語的複雑さ、語彙的複雑さ、正確さ、流暢さ）を独立変数として、決定木分析を行った。その結果、以下の図2に示すような決定木が得られた。



注. FL = 流暢さ, AC = 正確さ, LC = 語彙的複雑さ

図2 教師による最終評価を従属変数、客観的指標を独立変数とした決定木分析の結果

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、71.79を基に2つの主要な分岐が形成されている。教師は、「流暢さ」のスコアが71.79以上であった生徒のパフォーマンスには、最も高い評価である3を付けていることが分かる。ここでも、上記の重回帰分析の結果と同様に、「流暢さ」が教師による生徒のやり取りの評価において重要な観点であることが示された。

「流暢さ」のスコアが71.79より低い生徒のパフォーマンスは、「正確さ」の80.56を閾値に分岐している。この閾値より高い生徒のパフォーマンスには、2.5を平均に、2～3の範囲で総合的評価をしていることが示されている。最後に、80.56を閾値より低い生徒のパフォーマンスについては、83.50を閾値として「語彙的複雑さ」を基準に評価が行われている。非常に興味深い点として、語彙的に複雑であると評価が下がることが分かる。この点は、上記の担当教師のコメントにあげられていた普段の指導と、それに基づく評価方針と一致している。つまり、生徒同士で会話する際には、伝わることを重要視しているため、必要以上に難しい語彙を使うことは高く評価されないことが示された結果であると考えられる。

以上のことから、CAFによる客観データが教師の評価を支援する可能性について、実際に授業と評価を行った中学校教員との議論を通じて、今後の方向性を検討した。その中で挙げられた意見は、以下の3点に集約される。まず、分析にかかる労力の大きさに関する懸念が挙げられた。生徒のパフォーマンスを録画したビデオファイルをすべて書き起こし、ASユニットに分割し、WPMや統語的・語彙的複雑さを算出する作業には、予想を上回る労力が必要であった。このように多大な作業を大学に依頼するのは附属中学校でさえも現実的に困難であるという見解が多数を占めた。また、労力の大きさに加えて、生徒へのフィードバックに要する時間の長さも指摘された。実際、今回の調査では2クラス68名分のデータを処理したが、生徒に結果をフィードバックするまでに約1カ月を要した。このような長期間の遅延がある場合、生徒本人が1カ月前に行ったパフォーマンステストの内容を正確に記憶していることを期待するのは難しいと考えられる。

これらの懸念、特に分析にかかる労力の大きさについては、テクノロジーの進歩によって今後の改善が期待される。近年、AIやコンピューター技術は著しい発展を遂げており、音声の書き起こしや書き起こされたテキストの分析精度は飛躍的に向上している。このような技術環境が整えば、音声ファイルを読み込むだけで今回採用したような客観的指標を瞬時に算出できるようになる可能性は十分に考えられる。

次に挙げられた意見として、客観的指標の数値が生徒にとって分かりにくい点への懸念がある。例えば、「WPMが80」というフィードバックだけで、その意味や意義を明確に示さなければ、生徒にとって評価結果の解釈が難しいという指摘がなされた。具体的には、英語の母語話者が日常会話で達成する数値や、日本人学習者、特に中学生のような初級学習者が目指すべき基準を明示する必要があるとされた。このような情報がなければ、到達度評価や診断的評価としての役割を十分に果たすことができないとの懸念が示された。その結果、今回の評価では、生徒が関心を持って参照したのは主に教師による最終評価であり、CAFの指標そのものに注目する生徒は少なかったという報告があった。

この点については、健康診断における各種数値と同様に、CAFにおいても基準値を併せて提示する必要があることは事実である。しかし一方で、基準値を設定するには、話し言葉を客観的指標で分析する試みをさらに積み重ね、豊富なデータを収集する必要がある。特に、日本人中学生を対象とした研究は現時点で十分とは言えず、日本人中学生が話す標準的な英語の特徴や目標とすべき基準に関する知見はまだ限定的であるのが現状である。上述のように、現段階ではCAFの分析に多大な労力と時間を要することは否めないが、当該分野の研究を全国規模で実施し、さまざまな背景や特性を持つ生徒から得られるデータに基づいて、日本人中学生の「話す」パフォーマンスの実態を客観的指標で描写する研究を進めることが重要である。

最後に、CAFから得られる情報については、「教師の直観と同じ」であるものの、客観的なデータに基づく信頼性や説明責任の点を除けば、コストに見合う成果が得られにくいという指摘がなされた。特に、ユニットレベルの詳細な分析は中学段階では過剰に細分化され、指導に活用しにくい情報に留まる可能性があるとの懸念が挙げられた。たとえば、小さな差異であっても虫眼鏡や電子顕微鏡を用いれば観察可能であるが、それらの微細な差が将来的に大きな差異を生むのであれば、こうした精密な分析は妥当と言える。しかし、単に差異を見つけて評価結果として提示すること自体が目的化するのであれば、教育的視点からは本末転倒であるという批判がなされた。実際、本実践で採用したCAFの基準もさらに細分化することは可能であり、CAF以外にも多様な指標が提案されている。研究の場においては、こうした指標を算出して授業中の生徒の活動など通常の観察では捉えられない微細な点を議論することには一定の価値があるかもしれない。しかし、学校現場での実践に結びつけた取り組みとしては、評価の目的に即した妥当な指標を選択し、指標の数を可能な限り絞ってフィードバックを行うなど、実現可能性（practicality）を重視した評価の在り方が求められる。

結論として、本実践での教師による主観的評価とCAFに基づく客観的データの比較については、パフォーマンス評価のたびに生徒へフィードバックするというよりも、年間数回に限定して実施し、教師自身が自己の評価傾向を振り返る手段として活用するのが現実的であるとの結論に至った。実際に授業と評価を担当した中学校教員は、CAF分析により「流暢さ」と「正確さ」が有意な独立変数として抽出された一方で、「統語的複雑さ」と「語彙的複雑さ」が有意ではなかったという結果を受け、自身の評価が誤っていなかったという確信を得るとともに、評価に対する不安を大幅に払拭できたと報告している。

4. まとめと課題

本研究では、中学校の英語授業における生徒のパフォーマンステストのパフォーマンスを、教師の主観的評価と客観的指標（CAF）を組み合わせた評価手法の妥当性と信頼性について検討した。参加者は、国立大学教育学部附属中学校の68名の生徒で、ペアで「学級担任に何をプレゼントするか」というテーマに基づき話し合うパフォーマンスを行った。パフォーマンスは「流暢さ」「正確さ」「語彙的複雑さ」「統語的複雑さ」の4つの観点に基づいた数値データを大学で算出し、それを教師の観察による評価と組み合わせて最終評価を行った。教師による最終評価に対する4つの客観的指標（統語的複雑さ、語彙的複雑さ、正確さ、流暢さ）の影響を調べるため、重回帰分析を実施した結果、最終評価に有意な影響を与えたのは「正確さ」と「流暢さ」であり、それぞれの標準化回帰係数は0.33と0.50であった。一方、「統語的複雑さ」と「語彙的複雑さ」

の影響はほとんどなく、複雑な英語の使用が評価に大きく影響しないことが示された。決定係数 R^2 は0.37であり、教師の最終評価の37%が4つの客観的指標で説明でき、残りの63%は他の要因によって決定されることが分かった。

決定木分析の結果、「流暢さ」が教師による評価の最も重要な基準であり、流暢さのスコアが71.79以上の生徒は最も高い評価を得ていることが示された。また、「流暢さ」のスコアが低い生徒は、「正確さ」を基準に評価が分岐し、「正確さ」のスコアに基づいて評価がされていることが分かった。さらに、「語彙的複雑さ」が高い生徒ほど評価が低くなる傾向も観察され、教師が「伝わること」を重視し、難しい語彙を使うことを必ずしも高く評価しない方針を反映していると考えられる。

CAFを使用した分析に関しては、いくつかの課題も明らかになった。特に、パフォーマンスの録画をすべて書き起こし、WPMや語彙的・統語的複雑さを算出する作業には予想以上の労力が必要であり、実際の学校現場での実施には非現実的であるという意見が多く寄せられた。また、生徒へのフィードバックにかかる時間が長期間にわたるため、フィードバックを受けた際に、生徒が当時のパフォーマンスを正確に記憶していることが難しいという問題も指摘された。この点については、AI技術や音声分析技術の進展により、将来的にはより効率的な分析が可能になると期待される。多くの課題が示されたが、本実践報告が、これらをひとつひとつ検討する追実験のきっかけとなり、より妥当性と信頼性の高い中学生のスピーキング評価方法の確立に寄与することを期待する。

引用・参考文献

- Foster, P., Tonkyn, A. & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21, 354–375.
- Housen, A., & Kuiken, F. I. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, 30(4), 461–473. <https://doi.org/10.1093/applin/amp048>
- Najmi, K., Siyyari, M., Abassian, G. (2020). The Relationship Between Three Components of CAF in Learners' Oral Performance at Different Proficiency Levels. *Journal of English Language Pedagogy and Practice*, 13(26), 254–284. doi: 10.30495/jal.2020.683555
- 文部科学省. (2017). 『学習指導要領 中学校 [告示]』. 文部科学省. <https://www.mext.go.jp/>
- 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