第30回 研究助成

A 研究部門 ● 報告Ⅲ ● 英語能力テストに関する研究

RTWタスクにおけるEBBルーブリックの有用性

― 外部英語試験への架け橋 ―

研究者：福島県／福島大学大学院　在籍　久保田 恵佑

《研究助言者：大友 賢二》

概要　本研究は，技能統合的ライティングタスクの採点における妥当性や診断的機能を高める方法の検討を目的とし，TEAP（Test of English for Academic Purposes）のRTW（Reading-to-WriteTask）タスク（ライティングタスクB）向けのEBB（Empirically derived, Binary-choice, Boundary-definition scale）ルーブリックを作成し，多相ラッシュ分析を用いて得点化推論と波及効果推論について妥当性検証を行った。そして，TEAPルーブリック（Weir. 2014, pp. 16－20）と比較した際の診断的機能についても検証を行った。

その結果，本研究にて作成されたEBBルーブリックは，(1)得点化推論において，7個の前提のうち4個の証拠を提示できた，(2)波及効果推論において，2個の前提のうち2個の証拠を提示できた，(3)TEAPルーブリックに比べて診断的評価に適している，ということが示された。

本研究から，EBBルーブリックは，TEAPライティングタスクBにおいて，(1)採点の信頼性向上に寄与する，(2)採点のしやすさや採点結果の解釈のしやすさ向上に寄与する，(3)TEAPルーブリックに比べて診断的フィードバックを受験者に与えるのに適している可能性があることが示唆された。

1 はじめに

技能統合型テストとは，「聞く」，「読む」，「話す」，「書く」の4技能において，複数技能の統合的活用が求められるテストのこと（Plakans, 2013a）であり，講義内容や関連する複数の資料をまとめて論じたものを評価するなど，大学を中心としたアカデミックな場において従来から行われており，その重要性は広く認識されている。また，ATC21S（Assessment and Teaching of Twenty-First Century Skills Project）によって提唱されている「21世紀型スキル」やOECDによる「キー・コンピテンシー」などが重要視されており，技能統合的評価は注目を浴びてきている（松本，2016）。

日本においても，学習指導要領の改訂に伴い，「聞く」，「読む」，「話す」，「書く」を結びつけた統合的な言語活動を通したコミュニケーション能力育成の重要性がより一層高まっている（文部科学省，2017a，2018）。特に，高等学校学習指導要領においては，「論理・表現」が新設されることが決まり，資料を活用して文章を書くといった統合的な活動がより重視されている。また，大学入学者選抜改革が進められる中で，「聞く」「読む」「話す」「書く」の4技能を適切に評価するために，英語の外部検定試験（e.g.，英検，GTEC，TEAP）の活用が明記され（文部科学省，2017b），教育現場においてパフォーマンステストの重要性が今後さらに高まることが想定される。加えて，Green（2014）やSawaki（2017）は，日本の高校生，高校教員，大学生，大学関係者が4技能評価の導入に肯定的であることを報告しており，パフォーマンス評価としての技能統合型テストに対する関心

が高まりつつある。

日本で実施されている技能統合型テストの例を挙げるとTEAPがある。TEAPでは，課題文と図表の情報を統合・要約した上で結論を述べる，読んで書く（Reading-to-write, RTW）技能統合型ライティングタスクが設けられている（日本英語検定協会，n.d.）。こうしたタスクは，独立型ライティングタスク（independent writing task）よりもアカデミックな領域における真正性（authenticity）が高く（Cumming, 2013; Plakans, 2009, 2012, 2015），学習内容を現実世界に活かしやすいという，プラスの波及効果（テストによる影響）を期待できる。

しかし同時に，技能統合型テストには課題も多い。技能統合型評価の課題をまとめたCumming（2013）によれば，その課題の1つに，外部検定試験のようなハイステイクスな決定に関連する評価と学習や指導に関連する診断的な評価（diagnostic assessment）が混用されている，ということを挙げている。このCummingの指摘を日本の文脈で考えると，大学入試改革が進む中で，ハイステイクス（high-stakes）な試験を意識した総括的評価（summative assessment）がクラスルームでも重視されていくことが予想される。その一方で，日々の指導や学習を改善していくことを目的とする形成的評価（formative assessment），いわゆる「学習のための評価」[1]の重要性が軽視されてしまう恐れがある，ということが考えられる。クラスルームにおける技能統合的活動の指導をより一層充実させていくためには，フィードバックの利用しやすさ（accessibility）の質が高く，教師と学習者の双方に有益な情報を提供する，診断的機能の高い評価方法の検討が必須である。

日本においてこうした方法論を確立するためには，以下の2点に注意すべきである。1つは，多くの外部検定試験は熟達度の測定を目的としているという点である。熟達度測定テストを指導のためにそのまま導入することは，診断的・形成的評価も重視されるクラスルームの目的にそぐわない恐れがあることである。診断的評価を目的としていないルーブリック[2]の使用は，採点結果から得られる情報の少なさから，テスト後の指導や学習に活用されにくく，その波及効果がマイナスになることが考えられる。もう1つは，日本人学習者の熟達度が相対的に低いという点である。日本の国公立高校3年生およそ70,000人と90,000人を対象にした文部科学省（2015, 2016）の調査によると，調査対象者の80％以上がCommon European Framework of Reference（CEFR; Council of Europe, 2001）のスピーキング能力とライティング能力の観点でA1レベルであった。TEAPといった外部試験が対象にしているレベルが，主にA1からB2であることを考慮すると，既存のルーブリックの導入は学習者のレベルに適していない可能性がある。Fulcher（2003）は，こうした批判に対して，ルーブリックはテスト受験者のパフォーマンスに基づいたものにすべきでありと述べており，実証的に作成されたルーブリック，つまり，データに基づいて作成されたルーブリックの使用を推奨している。技能統合的タスク用のルーブリックがまだ確立されていない現状（Cumming, 2013）も考慮すると，実証的に作成されたルーブリックを導入することは，診断的機能を重視した評価の有効な方法の1つとして考えられる。

ルーブリックの作成には，その妥当性（validity）や信頼性（reliability），実用性（practicality）といった要素を熟慮する必要がある。Bachman and Palmer（1996）は，上記の要素に関して，作成者は各コンテクストに合わせた，上記3要素の適切なバランスを検討する必要がある，と述べている。クラスルーム評価というコンテクストを考慮すると，実施・採点・解釈の行いやすさといった実用性の要素は，重要な観点であると考えられる。ライティングをはじめとして，パフォーマンス評価は時間を要する。継続的に技能統合的タスクを実施していくために，採点に要する負担を軽減させる方法や結果の解釈を容易にする方法の検討が不可欠である。

そこで本研究では，まず，RTW（技能統合型ライティング）タスクの評価における課題と診断的評価における既存のルーブリックの問題点を整理する。そして，既存のルーブリックを導入する方法に代え，テスト受験者のパフォーマンスサンプルから実証的にルーブリックを作成する方法を導入し，その妥当性検証（validation）を行う。

2 先行研究

2.1 技能統合型ライティングタスクの定義と採点における課題

これまで，技能統合的ライティングタスクに関する研究は数多くなされており(e.g., Ascención, 2008; Cumming et al., 2005; Gebril & Plakans, 2013, 2014; Knoch & Sitajalabhorn, 2013; Plakans, 2009, 2012, 2015)，このタスクを定義づけようとする試みがなされている。例えば，Ascención(2008)は，RTWタスクを対象にして，その定義を「さまざまな教育目的のために，読むことと書くことを組み合わせた教育的タスク」(Ascención, 2008, p. 140)と述べている。また，Plakans(2012)は，技能統合型ライティングタスクを「タスク完遂のために2つ以上の技能を要するタスク」(Plakans, 2012, p. 249)と広く定義している。しかし，Cumming(2013)は技能統合型ライティングタスクには未だに明確な定義がないことを指摘している。その中で，Knoch and Sitajalabhorn (2013)は，これまでの技能統合型ライティングタスクに関する先行研究をレビューし，技能統合型ライティングタスクに求められる要素を以下のように定義することを提案している(Knoch & Sitajalabhorn, 2013, p. 306)。

(1)資料テクストからアイディアを得ようとすること
(2)アイディアを選定すること
(3)1つ，または複数の資料テクストからアイディアを統合すること
(4)インプット資料で使用されている言語を変形すること
(5)アイディアを整理・構成すること
(6)アイディアを関連づける資料を引用するといった表記上の習慣を用いること

これらの先行研究を整理すると，技能統合型ライティングタスクは，読むことや聞くことを，書くことと単純に組み合わせたものではなく，読むことや聞くことを通して得た情報を選定する，統合する，整理・構成するというプロセス[3]を経たアイディアを書くことが求められているタスクだといえる。技能統合型ライティングタスクの採点を行う際は，従来のライティングタスクの採点に必要な要素(e.g., 文法，語彙)に加え，技能統合的タスク固有の特徴も考慮する必要があるだろう。

しかしながら，技能統合型タスクは，上述のように独立型タスクよりも求められる要素が多く，そのプロセスが複雑であり(Brown, Iwashita & McNamara, 2005)，パフォーマンスの弁別的特徴にも影響を与えている。技能統合的ライティングタスクにおけるパフォーマンスの特徴の違いを検討した研究は複数あるが(e.g., Cumming et al., 2005, Gebril & Plakans, 2013)，その特徴を熟達度に応じて弁別する閾値(threshold)に未だ明確なものはなく，採点における困難さは，今後も検討されるべき課題である(Cumming, 2013)。

技能統合型ライティングタスクの採点における困難点の代表的なものは，インプット資料に関する評価である。例えば，Gebril and Plakans (2014)は，思考表出法を用いて，採点者が引用の形式や資料使用の適切さに関する判断に困難を感じていることを明らかにし，採点者によって重要と判断する部分が異なる傾向があることを示した。採点時の困難さは，採点の妥当性・信頼性・実用性に影響を及ぼす可能性がある。それゆえ，クラスルームで技能統合的ライティングタスクを実施する際には，可能な限り採点の困難さを軽減したルーブリックを考える必要がある。Knoch(2011)は，「ルーブリックは，ライティング評価における事実上のテスト構成概念として機能する」(Knoch, 2011, p. 81)と述べており，技能統合的ライティングタスクのルーブリックに関する実証的な検証が，それぞれの目的に応じて実施される必要がある。

2.2 ルーブリックの種類と作成方法

Weigle(2002)は，主なルーブリックの分類として「総合的ルーブリック(holistic rating scale)」，「分析的ルーブリック(analytic rating scale)」を挙げ，それぞれの特徴を信頼性，構成

概念妥当性(construct validity), 実用性の観点からまとめている。Weigleによると, 総合的ルーブリックは1つの観点から採点を行うことから採点が比較的早く, 実用性が高いとされている。一方で, 分析的ルーブリックは観点を分けて採点を行うため, 採点に時間を要する。しかしながら, 診断的機能の観点では, 総合的ルーブリックよりも分析的ルーブリックの方が, その観点別という特徴から, より詳細な診断的情報を得られる(Knoch, 2009, 2011)とされている。

ルーブリックの作成には主に2つの方法がある。1つは専門家の経験や知識から直感的, 経験的にルーブリックを作成する方法(measurement-driven method)であり, もう1つは学習者のパフォーマンスをもとに実証的に作成する方法(performance-based method)である(Fulcher, Davidson, & Kemp, 2011)。前者の方法は, a priori な方法とされており(Fulcher, 1996), 能力記述文が曖昧であることが多い, 学習者のパフォーマンスに見られない能力記述文が多く含まれる, との批判を受けている(Fulcher, 1996; Turner & Upshur, 2002; Upshur & Turner, 1995)。こうしたルーブリックは, その能力記述文の性質から, パフォーマンスの弁別力を低下させ, その結果, 重要な診断的情報を失わせる恐れがあることが指摘されている(Knoch, 2009)。後者の方法は, 学習者からパフォーマンスサンプルを集め, ルーブリック作成者による協議や談話分析によって, サンプルの特徴を弁別し, ルーブリックを作成する方法である(Fulcher et al., 2011)。この方法は, 前者の直感や経験に基づいたルーブリックが各教育環境における目的に合わないといった批判から提案され(Fulcher, 1996; Upshur & Turner, 1995; Turner & Upshur, 2002), ある特定の目的やコンテクストに応じて用いられるものである。Knoch(2009)は, この方法を受験者のパフォーマンスからルーブリックを作成するという特徴から診断的評価に適していると述べている。この方法を用いた代表的な例として, Upshur and Turner(1995)によって提案された「EBBルーブリック(Empirically derived, Binary-choice, Boundary-definition scale)」がある。

EBBルーブリックは,「実証的 (empirically) に作成され, 採点者による2択(binary)の選択を必要とし, スコアレベル間の境界(boundaries)を定義する」(Turner & Upshur, 1996, pp. 60-61)というものであり, パフォーマンスサンプルの顕著な特徴から作成された記述子が階層的に配置され, その記述子に対してYesかNoかを答えることで採点結果が得られる, という特徴がある(資料4)。これまで, ライティングタスクやスピーキングタスクを評価するために各研究で作成, 検証されており, その利点や欠点が明らかにされている(e.g., Fulcher, 2003; Hirai & Koizumi, 2013; Plakans, 2013b; Turner & Upshur, 1996, 2002; Upshur & Turner, 1995)。Fulcher(2003)は, EBBルーブリックはタスク特有型(task-specific)であり他のタスクへの一般化は難しいと指摘しているが, 同時に, この方法論の明瞭さと教育環境(pedagogic settings)における実用性の高さを評価している。Plakans(2013b)では, あるカリキュラムの中で, EBBルーブリック作成と使用を通して得られた知識や経験によって, 採点者が採点への自信をより深めた様子が報告されている。Turner and Upshur(2002)でも, Plakans同様, EBBルーブリックの作成が採点の信頼性に肯定的な影響与えることを報告されている。他にも, EBBルーブリックの使用に関する利点として, 信頼性の高さ(Hirai & Koizumi, 2013; Turner & Upshur, 1996)や使用と理解の容易さ(Turner & Upshur, 1996), Yes-Noの形式による採点者の記憶的負担の軽減(Fulcher et al., 2011)など, 採点者も学習者も享受できる利点が数多く示されている。

診断的評価の基本的な目的は, 学習者の強みや弱みを明確にすること, その後の学習や指導に活きる有益なフィードバックを与えることである(Lee, 2015)。EBBルーブリックは, パフォーマンスに基づいているという特徴から, 学習者のパフォーマンスをよりよく弁別し, その強みや弱みに関する診断的情報を得ることに適しているといえるだろう。一方で, 作成に関わる過程やYes-Noの形式から, 先行研究で述べられている技能統合的ライティングタスクの採点における困難さ(e.g., 引用の形式, 資料使用の適切さ)を軽減する可能性もある。EBBルーブリックの使用により, 学習者にも採点者にもプラスの波及効

果が生じる可能性があることは検証の価値があると考えられる。しかしながら，その検証の一方で，作成したルーブリックから得られるスコアが使用目的に対してどの程度適しているか，そのスコアの解釈はどの程度適切か，といった妥当性の側面も検証されなければならない。EBBルーブリックには，技能統合的ライティングタスクの採点の困難さを改善し，その妥当性を高めることに寄与する可能性については検討の余地がある。そこで次節では，妥当性を検証していく方法（妥当性検証）について概観する。

2.3 妥当性検証の枠組み

妥当性検証とは，妥当性[4]の要素がどの程度存在するのかを立証していくプロセスのことであり（小泉，2018），その実行可能な枠組みとしてよく用いられるのがKane（2006）による「論証に基づくアプローチ（argument-based approach）」である。このアプローチでは，2種類の論証に基づいて妥当性検証することを提案している。1つは，「解釈的論証（interpretive argument）」という，テストスコアを解釈し何かしらの決定のための推論（inference）や前提（assumption）を明示的に組み立てることである。もう1つは，「妥当性論証（validity argument）」という，解釈的論証の推論や前提を，具体的なデータ分析の結果を根拠に確認していくものである。

言語テストの分野においても，論証に基づく妥当性検証のアプローチが取り入れられている（e.g., Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Knoch & Chapelle, 2017）。例えば，採点プロセスを重視して作成されたKnoch and Chapelle（2017）の枠組みでは，得点化・一般化・説明・外挿・決定・波及効果という6つの推論（表1）とその推論に沿った理由，前提，証拠提示のための根拠（backing）が解釈的論証として設定されている。とりわけ，Knoch and Chapelleの枠組みは，テスト使用に関する側面を「決定」と「波及効果」の2つに分けており，テスト利用に関わる要因をより丁寧に扱うことができるとされている（小泉，2018）。本研究は，技能統合的ライティングの採点プロセス，特に，作成されたルーブリックによる観測得点の適切さの程度，テスト結果が指導や学習に与える波及効果の側面に焦点をあてているため，このKnoch and Chapelleによる枠組みを用いて，得点化と波及効果の側面（表2，表3）について妥当性検証を行う。

3 研究課題

本研究の目的は，RTWタスクに関する研究が積み重ねられている一方で，RTWタスク向けのルーブリックの有用性を高める研究やその診断的機能を重視した研究が少ないことを背景に，診断的評価により適したRTWタスク用のルーブリックを検討することである。そのために，EBBルーブリックを作成し，妥当性検証を行いながら，EBBルーブリックの有用性や診断的機能について検証していく。

手順は以下の通りである。まず，Knoch and Chapelle（2017）の枠組みに従って，採点の信頼性が主に関わる得点化推論，評価の診断的機能や

■表1：推論と対応する主張 (Knoch & Chapelle, 2017)

推論（Inference）	主張（Claim）
得点化（Evaluation）	テストパフォーマンスは，意図した特徴を持つ観測得点が得られる手順を用いて得点化されている
一般化（Generalization）	観測得点は，平行版であるタスク・テストフォーム間や，採点者間で一貫した値を示す期待得点の推定値である
説明（Explanation）	期待得点は，言語熟達度に関する構成概念によるものである
外挿（Extrapolation）	評価の構成概念は，目標言語使用領域での言語パフォーマンスの質を十分に説明している
決定（Decision）	パフォーマンスの質の推定値に基づいた決定は適切であり，よく伝えられている
波及効果（Consequence）	テストの結果は使用者にとって有益である

■表2: 得点化の推論に対する理由と前提,証拠提示のための根拠のまとめ (Knoch & Chapelle, 2017)

得点化の推論:テストパフォーマンスは意図した特徴を持つ観測得点が得られる手順を用いて得点化されている		
理由	前提	証拠提示のための根拠
A.ルーブリックの特性が開発者によって意図されたものである	1.分析的ルーブリックは仮説と同じように別々の能力を測っている	各ルーブリックの観点に関わる独自の因子を示す因子分析
	2.ルーブリック間の幅はそのレベルを弁別するのに十分である	ルーブリックの各レベル(ステップ)において十分な観測数を示す,多相ラッシュ分析;その他適切な量的検定や質的方法
	3.ルーブリックはテスト目的に求められるレベルに受験者を分けることができる	各ルーブリックの観点において十分な受験者分離を示す多相ラッシュ分析;2値的な尺度では二項・符号検定
B.採点者はタスクのレベルで高い信頼性を保って評価している	4.採点者はスコアレベル間でのパフォーマンスの違いを明らかにできる	採点者が異なるスコアレベルを使用していることを示す多相ラッシュ分析,その他の適切な質的検定(テスト状況に依存),採点者が全てのレベルにおいて採点に自信を持っていることを示す評価者の報告
	5.採点者はテストタスクに対してルーブリックを一貫して利用することができる	採点者の一貫性を示す統計分析(例.古典的テスト理論における信頼性のような技法や多相ラッシュ分析における平均平方値を用いる)
	6.採点者は記述子を容易に利用し,決定に自信を持っている	採点者のセルフレポート;インタビューや質問紙
	7.採点者はルーブリック(利用できる場合は観点も)を使用するのに徹底的,かつ定期的に訓練されている	専門家のレビュー; 採点者とテスト実施者へのインタビュー
	8.ルーブリックの例示を伴う十分な評価者支援の文書が利用できる	文書のレビュー; 採点者とテスト実施者へのインタビュー
	9.採点者はしかるべき資格を持っている	採点者雇用やその文書化に関する方針についての専門家のレビュー
	10.評価セッションは採点者パフォーマンスを最適化するようにデザインされている	採点セッション手順のレビュー; 採点者とテスト実施者へのインタビュー
	11.検出可能な採点者の特徴によって,系統的で構成概念に無関連な分散が,テスト設計者によって設けられている容認可能なレベルを超えた採点に影響を及ぼすことはない	バイアス分析からの結果(例.多相ラッシュ分析)が採点に影響を与えない,測定可能な採点者特徴を明らかにする;採点者のverbal protocolが採点者の認知プロセスと一貫していることを示す
	12.(利用できる場合に)ある特定の観点に対する採点者バイアスのレベルがテスト開発者によって設けられた容認可能なレベルの範囲内である	バイアス分析(例.多相ラッシュ分析)が,ルーブリックの観点を同様の方法で利用していることを明らかにする
	13.タスクタイプやテスト状況に関する系統的なその他の側面に対して採点者が示すバイアスのレベルがテスト開発者によって設けられた容認可能なレベルの範囲内である	バイアス分析(例.多相ラッシュ分析)が,タスクタイプ間や他のテスト状況間で,同様の方法で評価していることを明らかにする

■表3: 波及効果の推論に対する理由と前提,証拠提示のための根拠のまとめ (Knoch & Chapelle, 2017)

波及効果の推論:テストの結果は使用者にとって有益である		
理由	前提	証拠提示のための根拠
採点の手順はテスト使用者の最大の利益を促すようにデザインされている	1.テスト使用者は将来の指導と学習への情報を得るためにルーブリックを解釈することができる	テスト使用者へのインタビュー
	2.ルーブリックは(そしてルーブリックに直接関連するあらゆるフィードバックは)指導にプラスの波及効果がある	教師へのインタビュー; 教室での観察
	3.ルーブリックは(そしてルーブリックに直接関連するあらゆるフィードバックは)学習にプラスの波及効果がある	教師やテスト受験者へのインタビュー
	4.ルーブリックは採点者にプラスの波及効果がある	評価者へのインタビュー

採点の実用性に関わる波及効果推論に焦点を当てて,本研究で作成したEBBルーブリックの妥当性検証を行う。そして,TEAPのルーブリックと今回作成したEBBルーブリックを診断的機能の観点から比較を行う。この比較においては,本研究が,実証的に作成されたルーブリックの診断的機能について多相ラッシュ分析(Many-Faceted Rasch Measurement)を用いて検証を行ったKnoch(2009)やStory Retelling Speaking Test(SRST)向けのEBBルーブリックの妥当性検証を行ったHirai and Koizumi (2013)の研究目的と類似していることを踏まえ,KnochとHirai and Koizumiの手続きを参考にする。

研究課題1:
得点化の推論において,今回のEBBルーブリックは妥当性に関する証拠をあげることができるか(表5)。

研究課題2:
波及効果の推論において,今回のEBBルーブリックは妥当性に関する証拠をあげることができるか(表6)。

研究課題3:
EBBルーブリックは,TEAPのルーブリックに比べ,診断的評価に適しているか(表7)。

4 研究方法

4.1 協力者

本研究の協力者は,日本の国立大学に通う日本人英語学習者28名であった。テスト協力者の専攻は言語学,教育学,政治学,経済学など様々であった。

採点は,EBBルーブリック作成に関わった4名に加え,英語教育学を専攻する大学院生1名,大学生5名を加えた10名(TEAPのルーブリックでは9名)で行った。先にEBBルーブリックを用いて採点を10名で行った。EBBルーブリックでの採点を終えた約2週間後に,同様の採点者でTEAPルーブリックを用いて採点を行った。その際,EBBルーブリック作成に関わった大学院生1名は参加できなかったため,TEAPのルーブリックを用いた採点は9名で行った。なお,本研究では,各ルーブリックの解釈のしやすさを比較するために,事前の採点者トレーニングを実施しなかった。

4.2 使用テスト

本研究では,TEAP見本問題1(旺文社, 2015)のライティングタスクBを採用した。TEAPのラ

イティングタスクBは，テクストやグラフなど複数の資料内容を要約し，約200字で資料内容に沿った意見文を書くことが求められるRTWタスクの1つである。

TEAPのライティングタスクは，本来，要約型のTASK Aと上述のTASK Bを合わせて70分で実施される。本研究では，TEAPの実施要綱に則ってTASK AとTASK Bを70分で行い，分析にはTASK Bのサンプルのみを用いた。

4.3 手順

4.3.1 ルーブリック

本研究は，診断的機能を重視したルーブリックの検討を目的としているため，Knoch (2009, 2011)にならい，分析的ルーブリックを採用し，2種類のルーブリックを用いた。1つはTEAPが公表しているルーブリックである。TEAPのルーブリックは，観点ごと(Main Ideas, Coherence, Cohesion, Lexical Range & Accuracy, Grammatical Range & Accuracy)に，CEFRに基づいた4つのレベル(B2, B1, A2, Below A2)が設けられている。本研究では，多相ラッシュ分析を行うために，ルーブリックのレベルを，B2を4点，B1を3点，A2を2点，Below A2を1点へと変更した。

もう1つのルーブリックはEBBルーブリックである。EBBルーブリックの作成は，Turner and Upshur(1996)のEBBルーブリック作成手順を参考に行われた。Turner and Upshurは，EBBルーブリックの作成に関わる人数に関して，「評価の目的をよく知る者4名から8名」(Turner & Upshur, 1996, p. 61)を奨励しており，この点を留意した。評価観点は，TEAPルーブリックと同様の観点(Main Ideas, Coherence, Cohesion, Lexical Range & Accuracy, Grammatical Range & Accuracy)で作成に取り組んだが，CoherenceとCohesionの観点では，パフォーマンスサンプルを弁別できる特徴を複数確認できなかったため，CoherenceとCohesionを合わせて，1つの観点としてEBBルーブリック作成を行った。[5]

4.3.2 採点方法

採点する順番の採点への影響を避けるために，カウンターバランスをとって採点を行った。Main Ideas, Coherence (&) Cohesion, Lexical Range & Accuracy, Grammatical Range & Accuracyの順番で採点を行った組と，Grammatical Range & Accuracy, Lexical Range & Accuracy, Coherence(&)Cohesion, Main Ideasの順番で採点を行った組に採点者を半分ずつ分けた。

4.3.3 アンケートとインタビュー

採点終了後，採点協力者8名には2つのルーブリックに関するアンケートを実施した。アンケートはHirai and Koizumi(2013)を参考に作成された。その内容は，(1)より時間を要したルーブリックはどちらか，(2)より採点しやすかったルーブリックはどちらか，(3)(1)，(2)の回答に対する理由の自由記述，とした。なお，EBBルーブリック作成に関わっていない採点者には，(1)の質問に回答する際にはEBBルーブリックの評価観点がTEAPルーブリックに比べて1つ少ないことに留意するように指示した。また，テスト参加者28名のうち，2名にルーブリックに関するインタビューを行った。インタビューは，半構造化面接

■表4: TEAPルーブリックとEBBルーブリックの評価観点 (criteria)とレベル数の比較

TEAP	レベル	EBB	レベル
Main Ideas(MI)	4	Main Ideas(MI)	6
Coherence(MR)	4	Coherence & Cohesion(CC)	6
Cohesion(CS)	4		
Lexical Range & Accuracy(LRA)	4	Lexical Range & Accuracy(LRA)	4
Grammatical Range & Accuracy(GRA)	4	Grammatical Range & Accuracy(GRA)	5

法で行われ，内容は主に，(1)ルーブリックはわかりやすいかどうか，(2)ルーブリックの記述子の内容を自身の学習に活用できそうかどうか，(3)2つのルーブリックについての感想，とした。

4.4 分析方法

本研究で作成されたEBBルーブリックは評価観点ごとのレベル数が異なるため，Facets（Version 3.80.4; Linacre, 2018）を用いて，Partial Creditモデル（PartialCredit Model）で多相ラッ

■表5: 本研究で妥当化を行う得点化の前提（研究課題1）

得点化の推論：テストパフォーマンスは，意図した特徴を持つ観測得点が得られる手順を用いて得点化されている		
論拠	前提（研究課題）	証拠提示のための情報源
A.ルーブリックの特性が開発者によって意図されたものである	2.ルーブリック間の幅はそのレベルを弁別するのに十分である［(1) EBBルーブリックのレベルは弁別に十分か］	・ルーブリックのstrataの値 ・Bond and Fox（2015）による4つの基準
	3.ルーブリックはテスト目的に求められるレベルに受験者を分けることができる［(2)EBBルーブリックは受験者を弁別するのに十分か］	・受験者のstrataの値
B.採点者はタスクのレベルで高い信頼性を保って評価している	4.採点者はスコアレベル間でのパフォーマンスの違いを明らかにできる（得点化の推論におけるA2に同じ）	・Bond and Fox（2015）による4つの基準
	5.採点者はテストタスクに対してルーブリックを一貫して利用することができる［(3)採点者はEBBルーブリックを一貫して利用することができているか］	・採点者のinfit平均平方値
	6.採点者は記述子を容易に利用し，決定に自信を持っている［(4)採点者は自信をもってEBBルーブリックを用いているか］	・採点者への質問紙調査における記述
	11.検出可能な採点者の特徴によって，系統的で構成概念に無関連な分散が，テスト設計者によって設けられている容認可能なレベルを超えた採点に影響を及ぼすことはない［(5)採点者と受験者，採点者と評価観点の間に評価バイアス傾向はみられるか］	・有意バイアスの割合
	12.（利用できる場合に）ある特定の観点に対する採点者バイアスのレベルがテスト開発者によって設けられた容認可能なレベルの範囲内である（得点化の推論におけるB11に同じ）	・有意バイアスの割合

■表6: 本研究で妥当化を行う波及効果の前提（研究課題2）

波及効果の推論：テストの結果は使用者にとって有益である		
論拠	前提（研究課題）	証拠提示のための情報源
採点の手順はテスト使用者の最大の利益を促すようにデザインされている	3.ルーブリックは（そしてルーブリックに直接関連するあらゆるフィードバックは）学習にプラスの波及効果がある［(1) 受験者はEBBルーブリックから診断的情報を得ることができるか］	・受験者へのインタビュー結果
	4.ルーブリックは採点者にプラスの波及効果がある［(2) 採点者はEBBルーブリックの利点を感じることができるか］	・採点者への質問紙調査における記述

シュ分析を行った。多相ラッシュ分析は，基本のラッシュモデルを拡張したもので，2つ以上の相（facet）を加え，採点者の行動や採点者の厳しさといった詳細な情報を同一の尺度上で解釈することを可能にした分析方法である（e.g., Eckes, 2011）。本研究では，この多相ラッシュ分析に受験者，採点者，評価観点（criteria）の3層を分析に含め，3つの研究課題に取り組んでいく。

ラッシュモデルへの適合度の基準は，本研究がルーブリックを扱ったものであるため，Bond and Fox（2015）によるrating scaleの基準0.60～1.40を採用した。0.60未満であればラッシュモデルの予測に一致しすぎているオーバーフィット（overfit），1.40より大きければラッシュモデルの予測に一致していないアンダーフィット（underfit）と解釈した。

また，各相をいくつのグループに分けるかの指標であるseparationとstrataの区別に関して，Linacre（2018）の区別方法を参考にした。Linacreは，「非常に高い，ないし低いスコアが，おそらく高い，ないし低い能力によって生じている」（Linacre, 2018, p. 328）場合はstrataを用いて結果を解釈すべきとしており，本研究でも，採点結果は受験者のライティング能力から生じたと考えられるため，strataの値を用いて解釈した。

4.4.1 研究課題1，研究課題2の分析方法

Knoch and Chapelle（2017）の枠組み（表2, 表3）に則って，得点化と波及効果の推論において，多相ラッシュ分析から妥当化できるもの（表5, 表6）を，研究課題に沿って挙げていく。

4.4.2 研究課題3の分析方法

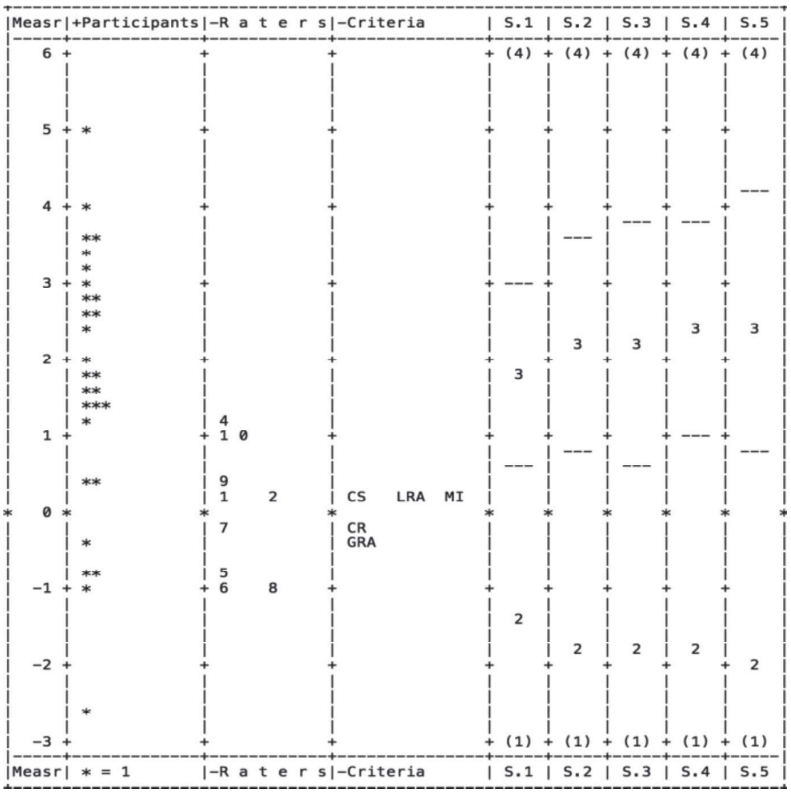
Knoch（2009）とHirai and Koizumi（2013）の多相ラッシュ分析の解釈方法を参考に，下記の観点で2つのルーブリックを比較する（表7）。

5 結果と考察

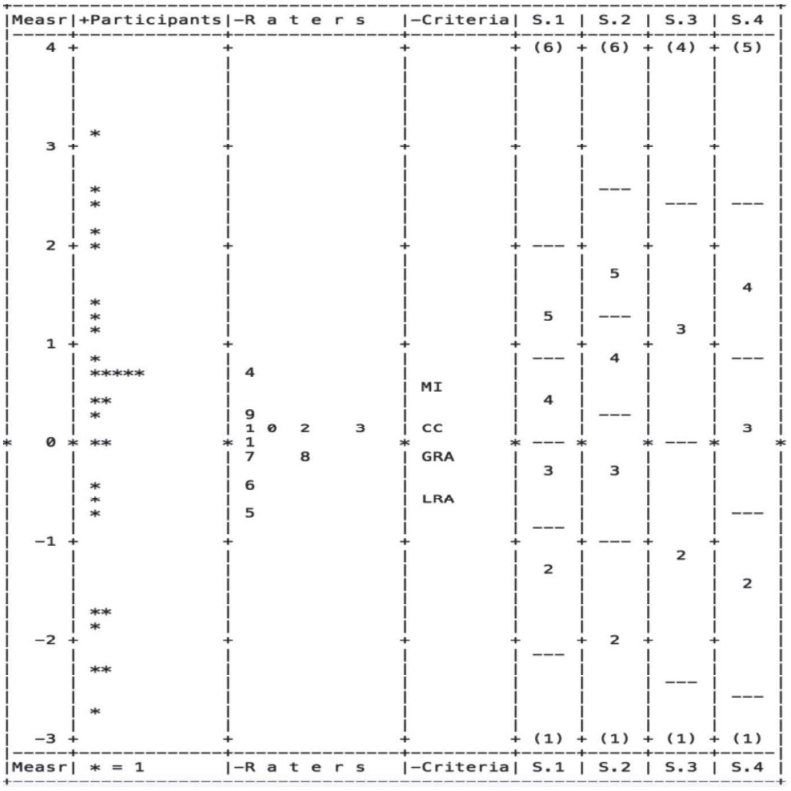
各変数マップ（図1, 図2）は，左から，ロジット（logit）という共通尺度で表した推定値（measure）と，テスト参加者，採点者，評価観点の3相を示している。推定値が上がる（下がる）に

■表7: ルーブリック比較の観点（研究課題3）

分析の観点	解釈	情報源
(1) ルーブリックの識別力 (Discrimination of the rating scale)	受験者の層に関する比率（participants strata ratio）は，ルーブリックの識別力の優れた指標となり，高い分離の比率であれば，そのルーブリックはより識別力がある。	・participant strata ratio
(2) 採点者の分離 (Rater separation)	よく機能するルーブリックは，甘さや厳しさの観点において評価者間の差が小さい。つまり，採点者の層に関する比率（rater strata ratio）が小さいルーブリックはより良く機能するとみなせる。	・rater strata ratio
(3) 採点者信頼性 (Rater reliability)	the rater point biserial correlation index（採点者が受験者をどの程度同じように評価したかに関する測定値）とpercentage of exact rater agreement（採点者が他の採点者と同じスコアを何回与えたかをパーセンテージで表したもの）の2種類の値が高ければ，より良く機能するルーブリックであることを示す。	・the rater point biserial correlation index ・percentage of exact rater agreement
(4) 採点の変動 (Variation in ratings)	より良く機能するルーブリックは，一貫性のない採点をする採点者や過剰に一貫している採点者を減らすと考えられる。設けられた範囲内にinfit 平均平方値があれば，採点者は一貫してルーブリックを使えており，ルーブリックはより良く機能している。	・採点者のinfit 平均平方値
(5) ルーブリックの特性 (Rating scale trait)	より良く機能するルーブリックは，以下の2つを満たしている。①推定値（Average Measures）が，スコアレベルが上がるにつれて上がっている。②各スコアレベルに少なくても10個の採点が含まれている。	・average measures ・counts used



■図1: TEAPルーブリックの変数マップ



■図2: EBBルーブリックの変数マップ

つれて，受験者の相では，能力が高く（低く）なることを，採点者や評価観点の相では採点が厳しく（甘く）なることを示している。そして，マップの右側にあるS.1 からS.5は各ルーブリックの評価観点を示している。TEAPルーブリックでは，Main Ideas（S.1），Coherence（S.2），Cohesion（S.3），Lexical Range & Accuracy（S.4），Grammatical Range & Accuracy（S.5）となっており，EBBルーブリックでは，Main Ideas（S.1），Coherence & Cohesion（S.2），Lexical Range & Accuracy（S.3），Grammatical Range & Accuracy（S.4）となっている。

■表8: 3相のルーブリックの記述統計（TEAPルーブリック）

	平均値（標準偏差）	最小値～最大値	範囲	分離（separation）	層（strata）	信頼性（reliability）
受験者	1.73 (1.71)	-2.59~5.03	7.62	4.96	6.95	0.96
採点者	0.00 (0.76)	-0.99~1.13	2.12	3.50	5.00	0.92
評価観点	0.00 (0.25)	-0.43~0.22	0.61	1.48	2.31	0.65

■表9: 3相の記述統計（EBBルーブリック）

	平均値（標準偏差）	最小値～最大値	範囲	分離（separation）	層（strata）	信頼性（reliability）
受験者	0.26 (1.52)	-2.67~3.08	5.75	6.39	8.85	0.98
採点者	0.00 (0.39)	-0.78~0.71	1.49	2.80	4.06	0.89
評価観点	0.00 (0.40)	-0.53~0.57	1.10	4.70	6.60	0.96

5.1 研究課題1

(1) EBBルーブリックのレベルは弁別に十分か（得点化の前提A2に対応，B4に対応）

表9から，EBBルーブリックのstrataの値を確認すると6.60となっている。本研究のEBBルーブリックは，Main Ideasが6つ，Coherence & Cohesionが6つ，Lexical Range & Accuracyが5つ，Grammatical Range & Accuracyが4つのスコアレベルであることを考慮すると，パフォーマンスサンプルの弁別力の数値としては概ね適切であると考える。加えて，Bond and Fox（2015）の満たすべき特質（properties），以下の4つ（a～d）に則ってEBBルーブリックの診断を行っていく。

a）.閾値（threshold）の推定値が，スコアレベルが上がるにつれて上がっている。

表10から，Lexical Range & AccuracyとGrammatical Range & Accuracyの観点においては，スコアレベルとともに閾値の推定値が上がっていることが確認できた。一方で，Main IdeasとCoherence & Cohesionの観点においては，レベル4とレベル5の間に逆転が生じており，このスコア間の記述子については改善が必要である。

実際に，レベル4とレベル5を決定する記述子を検討してみる（資料4）。Main Ideasの観点では，「一部本文に基づいていない理由であるが，理由とともに結論を述べることができている」という記述子に対してYesと答えるとレベル5，Noと答えるとレベル4となる。ここで考えられるレベル間の逆転の原因は，Gebril and Plakans（2014）が述べるような，結論に付された理由が資料内容に基づいているかの評価が困難であったことではないかと考えられる。パフォーマンスに基づいて作成されたEBBルーブリックにおいても，先行研究にて述べられている資料内容の評価については課題が残った。

また，Coherence & Cohesionでは，「代名詞の照応関係がわかりやすい」という記述子に対し

てNoと答えるとレベル4となり，Yesと答えると「因果関係を表すディスコースマーカーを適切に使うことができている」の記述子に辿りつき，Noの回答でレベル5となる。ここで考えられるレベル間の逆転の原因として考えられるのが，「代名詞の照応関係がわかりやすい」という記述に関する採点者による認識の差である。事実，採点者アンケートの中に，「照応関係などの例が欲しかった」という記述が確認された。この点に関連した採点者トレーニングを行った上で，再検証することが今後求められる。

b). 閾値の推定値間の距離が，1.4以上5.0未満である。

表10から，Main Ideasの観点では，レベル2と3，3と4，4と5，5と6の間の距離（推定値の差）が1.4に満たなかった。Coherence & Cohesionの観点では，レベル4と5の間の距離が1.4に満たなかった。Grammatical Range & Accuracyの観点では，レベル4と5の間の距離が1.4に満たなかった。以上から，Main Ideasの観点は各レベルにおいて，記述子の改善，レベルの統合，採点者が必要であると考えられる。他の観点においても，一部基準に満たない部分があったため，同様の改善策が求められるだろう。しかしながら，本研究のEBBルーブリックは診断的機能を重視し，弁別力を高めることを目的としたため，推定値間の距離が近くなることは避けられないかもしれない。この点に関しても，a）の特質の場合と同様の改善策を講じた上で再検証することが望ましい。

c). 確率曲線 (probability curve) が，グラフ上にはっきりとした頂上がある。

資料1から，Lexical Range & Accuracyでは，それぞれのレベルに頂上が確認できた。Grammatical Range & Accuracyでは，レベル4の頂上が低く，平坦であるが，頂上は確認できた。一方で，Main IdeasとCoherence & Cohesionの観点においては，レベル4の頂上が確認できず，各レベルの頂上も明らかであるとはいえない。この点に関しても，上述の改善策を講じた上で再検討してい

■表10: EBBルーブリックの観点ごとの敷居 (threshold) 推定値

	Main Ideas	Coherence & Cohesion	Lexical Range & Accuracy	Grammatical Range & Accuracy
	Thresholds Measures			
レベル1				
レベル2	-1.45	-2.79	-2.19	-2.09
レベル3	-0.84	-1.04	0.70	-0.71
レベル4	0.42	0.92	2.12	1.00
レベル5	0.34	0.79		1.80
レベル6	1.53	2.11		

■表11: 適合度統計量の割合 (EBBルーブリック)

	値 < 0.60 (overfit)	0.60 ≦ 値 ≦ 1.40 (fit)	1.40 < 値 (underfit)
受験者	0.00	100.00	0.00
採点者	0.00	100.00	0.00
評価観点	0.00	100.00	0.00

く必要がある。

d).適合度統計値 (fit statistics)
が2.0以下である。

表11における採点者infit平均平方値から，全ての観点において基準を満たしており，ラッシュモデルの予測に概ね一致していたと考えられる。

Bond and Fox(2015)は，上記4つの特質を満たさない場合は，上述したような改善策(e.g.，記述子の修正，レベルの統合，採点者トレーニング)が必要であるとしている。本研究の結果は，a)とd)の特質における基準を満たしていたが，b)とc)の特質における基準を満たしていなかった。以上から，(1)の前提に対する妥当化の証拠としては十分であるといいがたいだろう。今後，Bond and Foxが述べる改善策を講じた上でのさらなる検証を行う必要があるだろう。特に，Main IdeasとCoherence & Cohesionの観点においてはレベル4を決定する記述子の修正，レベル5との統合，採点者トレーニングといった改善策を実施した上での再検証が求められる。

(2)EBBルーブリック尺度は受験者を弁別するのに十分か(得点化の前提A3に対応)

表9のparticipant strataの値(8.85)から，受験者は約9つのグループに分かれていた。今回のEBBルーブリックのスコアレベルが4から6，評価観点が4つであることを考慮すると，意図したレベルよりも細かく分けられていたといえる。本研究の目的は，診断的機能を重視してルーブリックの弁別力を高めることであったため，このような結果になったと思われる。しかしながら，ルーブリックの作成には作成者の目的に沿っていることが重要であると考えられるため，この前提については，今後も改善が必要である。以上より，本研究の結果は，得点化A3の前提に対する妥当化の根拠としては不十分であると考えられる。

(3)採点者はEBBルーブリックを一貫して利用することができているか(得点化の前提B5に対応)

表11，表16(4)から，採点者infit平均平方値(0.62～1.36)が基準の範囲内であったので，EBBルーブリックを用いた採点は，ラッシュモデルによる予想の範囲内で一貫して行われていたと考えられる。

(4)採点者は自信をもってEBBルーブリックを用いているか(得点化の前提B6に対応)

採点者に対するアンケートの結果から(資料3)，アンケートに回答した8名の採点者のうち，全員が，TEAPルーブリックに比べてという制約付きではあるが，EBBルーブリックは使いやすかったと回答した。また，自由記述の結果から，多くの採点者が自信をもってEBBルーブリックを用いたであろうと考える。

(5)受験者と採点者，採点者と評価観点の間に評価バイアス傾向はみられるか
(得点化の前提B11，B12に対応)

Facetsのバイアス分析の結果から，Linacre(2018)の基準，t値が絶対値2.00を超えているかどうかを参考にし，t値が絶対値2.00を超えている場合は有意なバイアス傾向があるとして解釈した。

表12から，受験者と採点者の組み合わせにおいては，0.39%の有意なバイアス傾向が見られた。採点者とルーブリックの組み合わせにおいては，17.5%の有意バイアス傾向が見られた。前者は非常に小さい数値であるため問題にはならないと考えられる。一方，後者は数値としては比較的高いように思われる。この傾向に関して，Eckes(2005)では，採点者トレーニングを行った場合でも，採点者とルーブリックの組み合わせにおいて37.90%という高い割合の有意バイアス傾向が確認されていることから，この組み合わせの中では比較的低い数値である可能性がある。

また，EBBルーブリック作成者と非作成者に分けて，採点者とルーブリックの組み合わせにおける有意バイアス数について検討した。表13，表14の結果から，採点者とルーブリックの組み合わせにおける評価バイアスは，7個中6個がEBB非作成者から生じたものであった。この結果から，Upshur and Turner(2002)が述べるように，EBBルーブリックの作成に従事することが採点の一貫性に影響を与えた可能性があることが示唆された。同時に，EBBルーブリック非作成者に

おいて，相対的に評価バイアスが多いことから，EBBルーブリックを使用する上での採点者トレーニングの必要性も示唆された。

加えて，評価観点ごとに有意バイアス数の検討を行った。表15から，Grammatical Range & Accuracyの項目において，比較的多い有意バイアスが確認された。文法に関連する評価バイアスは先行研究においてもよく見られる傾向であり（e.g., Schaefer, 2008），文法指導が強調されている日本の英語教育環境が影響を与える可能性があるとされており（Matsuno, 2009），今後も検討すべき傾向であるだろう。しかしながら，その他の評価観点においては，有意バイアスはほとんどなく，全体的に問題のない範囲であるといえる。今後の研究では，上記の観点を考慮しながら，評価バイアスの詳細について検討していく必要があるだろう。

■表12: t値が絶対値2.00を超えたデータ数の割合（採点者全体）

	受験者 × 採点者	採点者 × 評価観点（EBB）
絶対値2.00を超えた%	0.39（1/256）	17.50（7/40）

■表13: t値が絶対値2.00を超えたデータ数の割合（EBBルーブリック作成者）

	採点者 × 評価観点（EBB）
絶対値2.00を超えた%	6.25（1/16）

■表14: t値が絶対値2.00を超えたデータ数の割合（EBBルーブリック非作成者）

	採点者 × 評価観点（EBB）
絶対値2.00を超えた%	25.00（6/24）

■表15: 評価観点ごとの有意バイアス数

採点者	Main Ideas	Coherence & Cohesion	Lexical Range & Accuracy	Grammatical Range & Accuracy
EBBルーブリック作成者	1	0	0	0
EBBルーブリック非作成者	0	1	0	4
合計	1/10	1/10	0/10	4/10

5.2 研究課題2

(1) 受験者はEBBルーブリックから診断的情報を得ることができるか（波及効果の前提3に対応）

テスト受験者2名へのインタビュー結果（資料2）から，この前提に関して確認していく。インタビューを行った2名の受験者からは，総じてEBBルーブリックに肯定的である様子がうかがわれた。特に，Fulcher et al.（2011）が述べるような，EBBルーブリックのYes-No形式による利点を採点者だけでなく，受験者も享受できた様子を確認できたことは，示唆に富むと思われる。この点に関して，インタビューのデータ数を増やすなどして，さらなる検証が望まれる。

また，EBBルーブリックのレイアウトの見やすさに関する言及も注目に値する。Lee（2015）は，スコアレポートの観点から，フィードバックにおける視覚的情報の重要性について述べている。EBBルーブリックの見やすさという視覚的効果に関連する点は，今後も検証していく価値がある

だろう。

総じて，受験者28名中の2名のみのインタビュー結果ではあるが，受験者はEBBルーブリックから自身の強み弱みに関する診断的情報を得ることができていたといえる。

(2) 採点者はEBBルーブリックの利点（e.g., 採点のしやすさ，記述子の理解のしやすさ）を感じることができるか（波及効果の前提4に対応）

採点者へのアンケート結果（資料3）から，注目すべきことは以下の3点であると思われる。1点目は，Yes-No形式に慣れない採点者も一部見られたが，多くの採点者がその形式による記憶的負担の軽減（Fulcher et al., 2011）や採点の実用性の高さ（Turner & Upshur, 1996）を利点として認識することができた，という点である。本研究においても，先行研究と同様の利点を採点者も認識することができていたといえるだろう。2点目は，記述子の具体性に関する点である。多様な要素を広く採点するTEAPルーブリックに比べ，EBBルーブリックの記述子は採点する要素が限定的であるため，具体性が高まり，その結果，採点への取り組みやすさに影響を与えたのだと考えられる。3点目は，EBBルーブリックのレイアウトへの記述がなされた点である。EBBルーブリックの実用性の高さはこれまでも示唆されてきたが（e.g., Hirai & Koizumi, 2013; Turner & Upshur, 1996），その根拠をEBBルーブリックのレイアウトによるものであるとした研究はこれまでなく，EBBルーブリックの有用性を高める新たな要素であると考えられる。もちろん，EBBルーブリックは，作成者によってレイアウトが多少異なるものの，Yes-No形式に変化はないため，共通する部分も多いだろうと考えられる。EBBルーブリックのレイアウトに関する言及については，今後も検証する価値があるだろう。

上記の結果をまとめると，採点者はEBBルーブリックの利点を感じ，享受することができたといえる。

5.3 研究課題3

研究のEBBルーブリックの診断的機能について，表7の(1)から(5)に基づいて論じていく。

(1) ルーブリックの識別力に関して，図1，図2の変数マップと表16の統計値から，TEAPルーブリックを用いた際の方が受験者の推定値の範囲は大きく（7.62），EBBルーブリックの方が小さかった（5.75）。一方で，表16から確認できるように，TEAPのparticipant strataの値は6.95，EBBは8.85となっており，EBBルーブリックの方がより高い弁別力をもつといえる。

(2) 採点者に関して，(1)と同様に図1，図2と表16から判断すると，TEAPルーブリックに比べ，EBBルーブリックの方が変数マップ上のばらつきも小さく，また，rater strataの値も多少ではあるが小さい。つまり，EBBルーブリックの方が，採点の厳しさの差が小さいルーブリックとして，より良く機能していると考えられる。

(3) 採点者の信頼性に関して，まず，the rater point biserial correlation indexを比較すると，表16から，EBBルーブリックの方が0.06高く，多少ではあるが高い信頼性を示している。一方，percentage of exact rater agreementでは，どちらのルーブリックもラッシュモデルの予測値よりも高い数値を示していた。しかしながら，TEAPルーブリックのスコアレベルがEBBルーブリックよりも少ないことを考慮すると，percentage of exact rater agreementにおいては，TEAPルーブリックの方が高い信頼性を示すとは一概にはいえないと思われる。この点に関しては，さらなる検証が求められる。

(4) 採点の変動に関して，表16から，TEAPルーブリックとEBBルーブリックどちらのルーブリックを用いた場合でもinfit平均平方値は基準の範囲内であった。つまり，採点者はいずれのルーブリックを用いても一貫して使え

ており，どちらのルーブリックもより良く機能しているといえる。

(5) ルーブリックの特性に関して，表17から，TEAPルーブリックを用いた場合，スコアレベルが上がるにつれて各観点の推定値が上がっていることが確認できた。一方で，全観点でレベル1の使用が10個に至らず，表7の(5)に対応する基準は満たされていなかった。また，表18から，EBBルーブリックを用いた場合も，スコアレベルが上がっていくにつれて各観点の推定値が上がっていることが確認できた。加えて，全観点で各レベルに10個以上の採点が確認でき，基準を満たしていた。以上から，各観点の各レベルが機能しているかという観点においては，EBBルーブリックの方がより良く機能しているといえる。

上記(1)～(5)の分析結果から，TEAPルーブリックとEBBルーブリックを診断的機能という観点から比較した場合，EBBルーブリックの方がより良く機能すると考える。

6 結論と今後の課題

本研究は，TEAPのRTWタスク向けのEBBルーブリックを作成し，多相ラッシュ分析を用いて，得点化推論と波及効果推論についての妥当性検証とEBBルーブリックの診断的機能についての検証を行った。本研究の研究課題に対する結果は以下の通りである。

(1) 得点化の推論において，7個の前提のうち4個の証拠を提示できた(研究課題1)
(2) 波及効果の推論の推論において，2個の前提のうち2個の証拠を提示できた(研究課題2)
(3) 本研究のEBBルーブリックは，TEAPルーブリックに比べて診断的評価に適している(研究課題3)

しかしながら，得点化推論の中には証拠を十分に提示できなかった前提が3個あり，ルーブリックを改善した上で再検証する必要がある。本研究で作成されたEBBルーブリックは，多相ラッシュ分析による診断の結果，Main IdeasとCoherence & Cohesionの観点でいくつかの問題点が確認された。その1つが，ルーブリックの観点(Main IdeasとCoherence & Cohesion)において，スコアレベル間の逆転がみられたことである。今後の検証では，多相ラッシュ分析の結果や採点者アンケートの結果から，記述子の改善やスコアレベルの統合，採点者トレーニングを行った上で，本研究のEBBルーブリックの妥当性をより高めていく必要があると考える。また，妥当性検証には，今回検証した得点化と波及効果の他にも，一般化・説明・外挿・決定の推論があるため，目的に応じて，さらに妥当性検証を進めていく必要があるといえる。

全体を総括して，本研究における示唆を述べる。本研究では，得点化推論と波及効果推論の側面について，TEAPのRTWタスク(ライティングタスクB)向けのEBBルーブリックの妥当性検証を進めることができた。そして，EBBルーブリックが，TEAPルーブリックと比較して，診断的評価に適していることを示した。具体的には，研究課題1から，EBBルーブリックは，採点者内の信頼性，つまり，採点の一貫性向上に寄与する可能性があることが示唆された。研究課題2から，採点者・受験者の双方にプラスの波及効果(e.g.，採点のしやすさ，採点結果の解釈のしやすさ)を与える可能性があることが示唆された。研究課題3から，EBBルーブリックは，TEAPルーブリックに比べ，よりテスト受験者の強みや弱みに関する具体的なフィードバックを得る目的に適している可能性があることが示唆された。

本研究では，集められたライティングのサンプル数が28，インタビューのサンプル数が2と少なく，研究結果の一般化には不十分である。今後の検証では，それぞれのデータサンプル数を増やして，本研究の結果が体系的なものであるかを確認し考察を深めていくことが期待される。加えて，本研究で対象とした受験者は，国立大学の大学生であった。本研究で使用したTEAPが高校生の大学入試への利用が想定されたテストであることを考慮すると，今後の研究では高校生を対象にして再検証する必要があるといえる。

■表16: 2つのルーブリック解釈に関する統計値

観点	指標	TEAP	EBB
(1)ルーブリックの識別力	participant strata	6.95	8.85
(2)採点者の分離	rater strata	5.00	4.06
(3)採点者信頼性	rater point biserial correlation	0.44	0.50
	percentage of exact rater agreement	52.10%	43.20%
(4)採点の変動	infit平均平方値	0.77~1.20	0.62~1.36
	criteria separation ratio	1.48	4.70

■表17: TEAPルーブリックの評価観点ごとの統計値

(5)ルーブリックの特徴	Main Ideas	Coherence	Cohesion	Lexical Range & Accuracy	Grammatical Range & Accuracy
	Counts used (%) [Average Measures]				
レベル1	5 (3%) [-2.69]	2 (1%) [2.61]	3 (2%) [-3.02]	2 (1%) [0.02]	1 (1%) [-1.32]
レベル2	59 (33%) [-0.08]	56 (31%) [0.11]	59 (33%) [-0.11]	70 (39%) [0.25]	49 (27%) [0.39]
レベル3	63 (35%) [1.91]	79 (44%) [2.28]	85 (47%) [1.99]	73 (41%) [2.00]	91 (51%) [2.30]
レベル4	53 (29%) [3.20]	43 (24%) [3.44]	33 (18%) [3.40]	35 (19%) [3.35]	39 (22%) [4.02]

■表18: EBBルーブリックの評価観点ごとの統計値

(5)ルーブリックの特性	Main Ideas	Coherence & Cohesion	Lexical Range & Accuracy	Grammatical Range & Accuracy
	Counts used (%) [Average Measures]			
レベル1	43 (19%) [-2.10]	14 (6%) [-1.98]	17 (7%) [-1.26]	18 (8%) [-1.45]
レベル2	38 (16%) [-1.09]	44 (19%) [-0.96]	62 (27%) [-0.29]	41 (18%) [-0.99]
レベル3	48 (21%) [-0.24]	69 (30%) [-0.18]	93 (40%) [1.01]	70 (30%) [0.49]
レベル4	32 (14%) [0.24]	37 (16%) [0.58]	60 (26%) [2.40]	56 (24%) [1.31]
レベル5	40 (17%) [0.69]	42 (18%) [1.29]		47 (20%) [1.85]
レベル6	31 (13%) [2.02]	26 (11%) [1.95]		

また，本研究で対象とした採点者が，EBB作成者が4名，非作成者が6名の合計10名と，小規模であること，採点者に関する調査方法がアンケートのみであったことも限界点として挙げられる。今後の検証では，採点者の数を増やし，インタビューも取り入れる必要があるといえる。

最後に，本研究で作成したEBBルーブリックは，診断的機能を重視したため分析的採点の形式を採用した。それゆえ，評価観点ごとのより詳細な情報を受験者も採点者も得られるものとなった。しかし同時に，分析的採点は全体的採点に比べ時間を要する（e.g., Weigle, 2002）ため，必ずしもすべての観点で採点する必要はないと思われる。各クラスルームの指導目的に応じて，評価観点を限定するといった工夫が求められるだろう。EBBルーブリックのような実証的なルーブリックの利点を最大限に享受するために，各クラスルームでの評価目的を明確にした上で，ルーブリックを作成し使用していくことが望まれる。

謝辞

本研究への機会とご支援をくださいました，公益財団法人 日本英語検定協会の皆様，ならびに選考委員の先生各位に厚く御礼申し上げます。特に，指導助言者である大友賢二先生には，大変有益なご助言を賜りました。深く感謝申し上げます。そして，福島大学の髙木修一先生には，本研究の実施から執筆に至るまで，お忙しい中たくさんの貴重なご教示と温かな励ましの言葉を頂きました。心より感謝申し上げます。また，大学院で共に学んだ，奥山拓弥さんにも多くのサポートを頂きました。最後に，データ収集にご協力いただいた大学生の皆様，採点にご協力いただいた大学生・大学院生の皆様に改めて御礼申し上げます。ありがとうございました。

注

(1) 近年，形成的評価や総括的評価に代わり，「学習のための評価」と「学習の評価」という用語が頻繁に用いられるようになっている（二宮，2015, p. 60）。

(2) 本研究では，“rating scale” に準ずる用語を広義に捉え，「ルーブリック（rubric）」で統一して使用している。

(3) 技能統合的ライティングタスクでは，アイディアの「整理・構成（organizing）」，「選定（selecting）」，「関連づけ（connecting）」といった下位プロセスを伴う，「談話統合（discourse synthesis）」（Spivey & King, 1989）という構成概念が生じることが発見されており（Plakans, 2009），この能力は，タスクの完遂に必要な能力であるとされている（Plakans, 2009; Yang & Plakans, 2012）。

(4) 妥当性とは，「テスト開発者（テスト作成者）がテストで図りたいと思う能力（構成概念）がどの程度測れているか，また，使用目的にどの程度合っているか」（小泉，2018, p. 38）を示すものであり，その要素は，「内容的要素（content aspect）」，「実質的要素（substantive aspect）」，「構造的要素（structural aspect）」，「一般化可能性的要素（generalizability aspect）」，「外的要素（external aspect）」，「結果的要素（consequential aspect）」に分けられる（e.g., Messick, 1996; 小泉，2018）。

(5) Plakans and Gebril（2017）は，「談話統合」の下位プロセスである，「整理・構成」の構成要素 ー「構成パターン（organizational patterns）」，「一貫性（coherence）」，「結束性（cohesion）」ーとライティングスコアの関連性を調査した。その結果，構成パターンと一貫性は，スコアに有意に関連しており，結束性は，有意に関連していないことが発見された。このことからも，本研究のルーブリックで，CoherenceとCohesionを統合して1つの評価観点としたことは問題ないと思われる。

参考文献（＊は引用文献）

＊ Ascención, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7,* 140–150. doi:10.1016/j.jeap.2008.04.001

＊ Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

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

＊ Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

＊ Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series, MS–29). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-05-05.pdf

＊ 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.

＊ Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press.

＊ Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly,* 10,1–8. doi:10.1080/15434303.2011.622016

＊ Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing,* 10, 5–43. doi:10.1016/j.asw.2005.02.001

＊ Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly,* 2, 197–221. doi:10.1207/s15434311laq0203_2

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

＊ Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing,* 13, 208–238. doi:10.1177/026553229601300205

＊ Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Education.

＊ Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing,* 28, 5–29. doi:10.1177/0265532209359514

＊ Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-writing tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10,* 9–27. doi:10.1080/15434303.2011.642040

＊ Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21,* 56–73. doi:10.1016/j.asw.2014.03.002

＊ Green, A. (2014). *The test of English for academic purposes (TEAP) impact study: Report 1—Preliminary questionnaires to Japanese high school students and teachers.* Tokyo: Eiken Foundation of Japan. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf

＊ Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly,* 10, 398–422. doi:10.1080/15434303.2013.824973

＊ Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

＊ Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26,* 275–304. doi:10.1177/0265532208101008

＊ Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?. *Assessing Writing, 16,* 81–96. doi:10.1016/j.asw.2011.02.003

＊ Knoch, U., & Chapelle, C. (in press).n Validation of rating processes within an argument-based framework. *Language Testing, 34,* 1–23. doi:10.1177/0265532217710049

＊ Knoch, U & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing, 18,* 300–308. doi:10.1016/j.asw.2013.09.003

＊ 小泉利恵. (2018).『英語4技能テストの選び方と使い方ー妥当性の観点からー』. 東京: アルク.

＊ 公益財団法人日本英語検定協会. (n.d.).「問題構成・見本問題」. Retrieved from http://www.eiken.or.jp/teap/construct/

＊ Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32,* 299–316. doi:10.1177/0265532214565387

＊ Linacre, J. M. (2018). *A user's guide to FACETS: Rasch-model computer programs Program manual 3. 80.* 4. Retrieved from http://www.winsteps.com/a/manuals.htm

＊ Linacre, J. M. (2018). FACETS: Rasch-measurement computer program (Version 3.80.4) [Computer software]. Chicago: MESA Press.

＊ 松本佳穂子. (2016).「技能統合的ライティングの評価」. 渡部良典・小泉利恵・飯村英樹・髙波幸代 (編).『日本言語テスト学会誌20周年記念特別号』. (pp. 128–131). 日本言語テスト学会. doi:10.20622/jltajournal.19.2_0

＊ Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing,* 26, 75–100. doi:10.1177/0265532208097337

＊ Messick, S. (1996). Validity and wash back in language testing. *Language Testing,* 13, 241–256. doi:10.1177/026553229601300302

＊ 文部科学省. (2015).「平成26年度英語教育改善のための英語力調査事業報告」Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1358258.htm

＊ 文部科学省. (2016).「平成27年度英語教育改善のための英語力調査事業報告」Retrieved from http://www.mext.

参考文献（＊は引用文献）

go.jp/a_menu/kokusai/gaikokugo/1377767.htm

＊文部科学省. (2017a).「中学校学習指導要領」Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/1384661.htm

＊文部科学省 (2017b).「高大接続改革の実施方針等の策定について（平成29年7月13日）」Retrieved from http://www.mext.go.jp/b_menu/houdou/29/07/1388131.htm

＊文部科学省. (2018).「高等学校学習指導要領」Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/1384661.htm

＊二宮衆一. (2015).「教育評価の機能」. 西岡加名恵・石井英真・田中耕治 (編).『新しい教育評価入門ー人を育てる評価のために』. (pp. 51-75). 東京: 有斐閣.

＊旺文社. (2015).『大学入試合格のためのTEAP実践問題集』. 東京: 旺文社.

＊Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26,* 561-587. doi:10.1177/0265532209340192

＊Plakans, L. (2012). Writing integrated items. In: Glenn Flucher and Fred Davidson (Eds.), *The Routledge handbook of language testing,* 249-261. New York: Routledge.

＊Plakans, L. (2013a). Assessment of Integrated Skills. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics, Vol. I* (pp.204-212). Hoboken, NJ: Wiley-Blackwell.

＊Plakans, L. (2013b). Writing scale development and use in a language program. *TESOL Journal, 4,* 151-163. doi:10.1002/tesj.66

＊Plakans, L. (2015). Integrated second language assessment: Why? What? How? *Language and Linguistic Compass, 9,* 159-167. doi:10.1111/lnc3.12124

＊Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing, 31,* 98-112. doi:10.1016/j.asw.2016.08.005

＊Sawaki, Y. (2017). University faculty members' perspectives on English language demands in content courses and a reform of university entrance examinations in Japan: A needs analysis. *Language Testing in Asia, 7*(13), 1-16. doi:10.1186/s40468-017-0043-2

＊Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25,* 463-492. doi:10.1177/0265532208094273

＊Spivey, N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly, 24,* 7-26. Retrieved from http://www.jstor.org/stable/748008

＊Turner, C. E., & Upshur, J. A. (1996). *Developing rating scales for the assessment of second language performance.* In G. Wigglesworth, & C. Elder (Eds.). The language testing cycle: From inception to washback (pp. 55-79). Australia: Applied Linguistics Association of Australia.

＊Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly,* 36, 49-70. doi:10.2307/3588360

＊Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12. doi:10.1093/elt/49.1.3

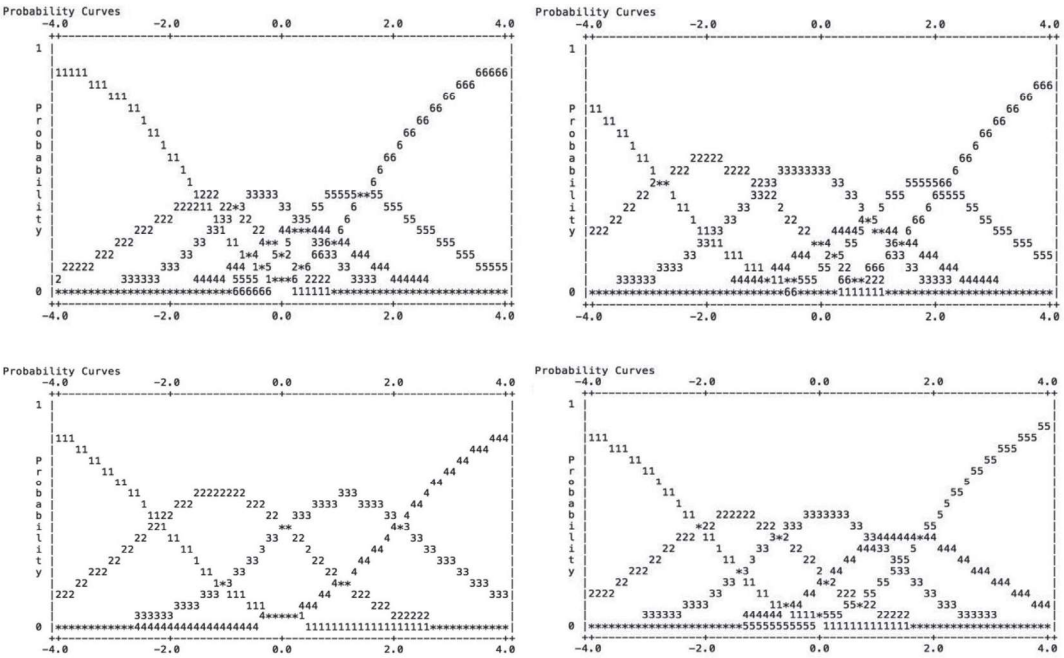
＊Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C.J. (2014). A Reserch Report on the Development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese University Entrances. Eiken Foundation of Japan.

＊Yang, H.-C. & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly,* 46, 80-103. doi:10.1002/tesq.6

資料1　EBBルーブリックの観点ごとの確率曲線

(左上:Main Ideas, 右上:Coherence & Cohesion, 左下:Lexical Range & Accuracy, 右下:Grammatical Raneg & Accuracy)



資料2　受験者へのインタビュー結果のまとめ(一部省略)

回答者	(1)ルーブリックはわかりやすいかどうか	(2)ルーブリックの記述子の内容を自身の学習に活用できそうかどうか	(3)2つのルーブリック(TEAPルーブリックとEBBルーブリック)についての感想
受験者1	EBBルーブリックは「〜できる」の書き方だからわかりやすいです。YesとNoで進んでいくのが良いと思います。	EBBルーブリックは具体的な説明が良かったです。それと,レイアウトが見やすくて,見る気になります。勉強に生かせそうなのは,EBBルーブリックだと思います。	EBBルーブリックは,自分の良くなかったところがわかるのは良かったです。TEAPルーブリックは,どのレベルに達しているかの判断ができる気がします。
受験者2	TEAPルーブリックもわかりやすかったし,EBBルーブリックもわかりやすかった。EBBルーブリックは,Yes-Noのおかげで,どこができないか,どこでできないかがわかった。	EBBルーブリックは,1つ1つ分かれているから理解しやすいし,具体的な言葉もわかりやすい。	TEAPルーブリックは,点数が目にすぐ入ってわかりやすい。

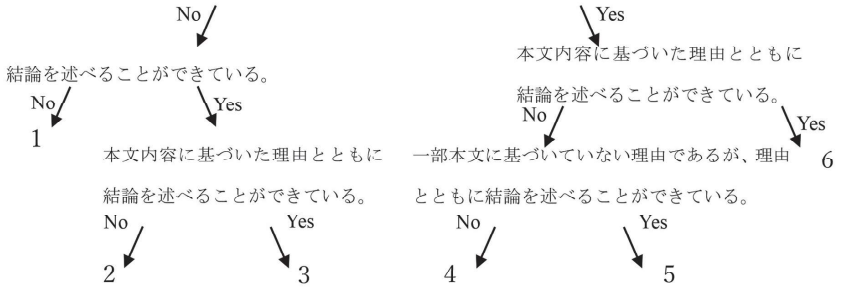
資料3　採点者アンケート結果のまとめ(一部省略)

採点者	質問1	質問2	質問3 (質問1に対して)	質問3 (質問2に対して)	質問4
Rater 2	EBB	EBB	TEAP（の採点基準は）実質3件法だったので採点が早かった。	EBBは尺度が細かいため、判断に時間がかかったものの、より自信をもって評価することができた。	無記入
Rater 4	TEAP	EBB	TEAPの場合は、記述子に複数の要素が含まれていたためレベルを見極めが難しい場合があり、迷ってしまったからだと考えられる。	EBBの場合は記述子がそれぞれ明確で、かつ、複数の要素にまたがっていることが少なかったため、レベルの判断がよりしやすかったように思う。	全体として、EBBの場合はサンプルを基に記述子を決定していることもあり、評価がしやすいと感じた。
Rater 5	EBB	EBB	EBBはTEAPオリジナルに比べて、評価基準の過程があらかじめ決められていたので、1つ1つ対応させていく作業に時間がかかったのだと思う。	評価基準の過程が決められていたという同じ理由から、EBBの方が迷わずに評価することができた。	EBBではCOHERENCEとCOHESIONが同じくくりになっていたので、文章の流れの観点は評価しやすかった。
Rater 6	TEAP	EBB	EBBの方が使いやすかったので所要時間はEBBの方がかかりませんでした。	EBBについては評価過程がYESかNOの選択の連続でしたので、TEAPよりも思考過程が明確になり、使いやすかったです。	EBBの場合の評価に至るまでの過程が明確になるところはひとつの強みだと感じました。
Rater 7	TEAP	EBB	EBBでは、YesかNoで進んでいき判断をしていくことができたため、TEAPよりも評価基準と合っているかを考えやすかったから。	EBBでも少し評価に時間がかかるものもあったが、評価基準が分かりやすく、やりやすいと思った。	TEAPよりも評価基準が細かく、フィードバックもしやすいのではないかと思った。
Rater 8	TEAP	EBB	TEAPの方が文章で評価基準が書かれてあり、EBBの方が使いやすかったから。	EBBの方がYesとNoで簡単に振り分けることができた。	簡単に行うことができたが、何個要点を書いてあったらYesにしてもいいのかが分からなく、戸惑ってしまった。
Rater 9	TEAP	EBB	TEAPはEBBに比べると評価観点の内容が具体的ではなかったため、悩ましい部分は考え直すことが多かったから。	評価の観点の内容がより具体的かで、ほとんど迷わず評価を進めることができたから。また、YESかNOの2択なので、どちらか片方を選ぶのはそこまで負担にならずに評価することができた。	TEAPにはなかった「重文・複文」の観点があり、最初はどのような観点で見ればよいか分かりませんでした。しかし、自分で調べ理解できた後は、負担なく評価を進めることができました。
Rater 10	同じ	EBB	一つ一つの観点ごとに評価していこうとすると何度も文章を読み返すこととなり、それはどちらの評価基準でも同じだったので、あまり時間に変わりがなかったのだと思います。	EBBの評価尺度の方では評価基準がYes/Noではっきりしているため、採点しやすかったです。また、樹形図のような形も含めて、見やすく、内容がすっきりしていたのが良かったのだと思います。	文字数が少なすぎるもの、設問文から直接書き写しているものなど、前提としてあまりよくないものに関する記述があると、より分かりやすかったかなと思います。

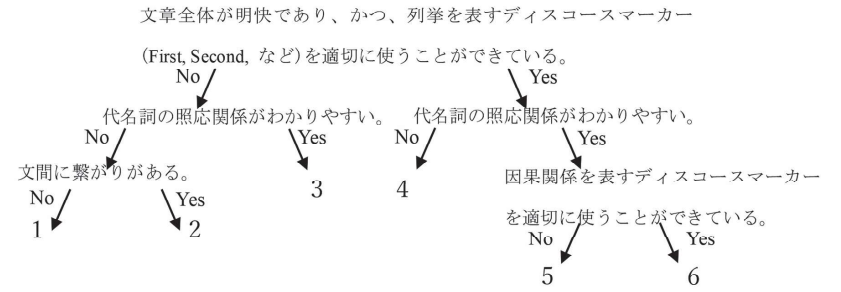
資料4　本研究で用いたEBBルーブリック

The EBB rating scale for TEAP Writing Task B

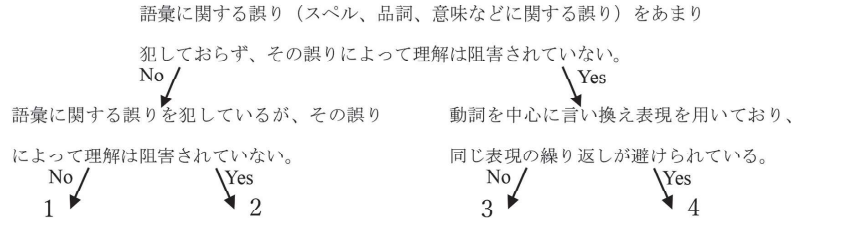
1. MAIN IDEAS



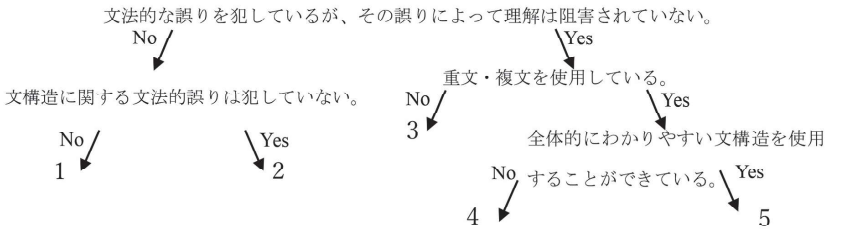
2. COHERENCE & COHESION



3. LEXICAL RANGE & ACCURACY



4. GRAMMATICAL RANGE & ACCURACY



資料5　採点者へのアンケート

評価者アンケート

・以下の質問に回答してください。（質問１、２は選択形式、質問３は自由記述）

1.　どちらの評価基準がより時間を必要としたか。
（選んだ記号を残し、その他の記号は削除してください）
ア．TEAP　　イ．EBB　　ウ．同じ

2.　どちらの評価基準がより使いやすかったか。
（選んだ記号を残し、その他の記号は削除してください）
ア．TEAP　　イ．EBB　　ウ．同じ

3.　質問１、質問２の回答に対する理由をそれぞれ述べてください。

4.　EBB rating scale について、改善点（記述子のわかりにくかった点など）や感じたことなど、自由に述べてください。

以上

ご協力ありがとうございました。