Shinshu University Journal of Educational Research and Practice, No.12, pp. 1-16, 2018
信州大学教育学部研究論集　第12号　pp. 1-16　2018年

＜学術論文＞

ライティング・パフォーマンス評価の検討

―含意尺度法，自己評価との相関分析，ラッシュ分析を用いて―

伊東哲　　**東京学芸大学大学院**
菊原健吾　　**信州大学大学院教育学研究科**
酒井英樹　　**信州大学学術研究院教育学系**

キーワード：ライティング，パフォーマンス評価，CEFR-J，ラッシュ分析

１．序論

1.1 本研究の目的

本研究[1]の目的は，CEFR 及び CEFR-J のレベルに階層性が想定されることに基づいて提案された，ライティング・パフォーマンス評価の方法の妥当性を検証することである。検証する評価方法とは，CEFR や CEFR-J のレベルを参考とした難易度の異なる複数のタスクを実施し，3 段階で評価することによって，学習者のパフォーマンスのレベルを特定する方法である。また，検証方法は，含意尺度による分析（Implicational Scaling, Hatch & Lazaraton, 1991）をおこなうとともに，CAN-DO リスト形式の自己評価との相関分析やラッシュ分析（Bond & Fox, 2001; 竹内・水本, 2012）を用いて算出した能力推定値との相関分析をおこなった。本節では，CEFR 及び CEFR-J について簡略に説明し，階層性を想定して提案された評価方法を解説する。

1.2 CEFR と CEFR-J

(1) CEFR とは

Common European Framework of Reference（CEFR）は，ヨーロッパの言語教育のシラバス，カリキュラムのガイドライン，試験，教科書等の向上のために一般的基盤を与えること，言語学習者が言語をコミュニケーションのために使用し，効果的に行動するために何を学び，どのような知識・技能を身につければ良いかを総合的に記述することを目的に欧州評議会によって開発された外国語能力の参照基準である（Council of Europe, 2001, p. 1）。また，CEFR は学習者の熟達度のレベルを明示的に記述し，それぞれの学習段階，また生涯を通して学習進度を測ることができるようにしている（Council of Europe, 2001; 投野編, 2013）。

CEFR の大きな特徴の一つとして，言語共通参照枠が示されていることが挙げられる。これは言語コミュニケーション能力を 6 段階で示しているものである。つまり，言語コミュニケーション能力を A（初級）「基礎段階の言語使用者」，B（中級）「自立した言語使用者」，C（上級）「熟達した言語使用者」の 3 つに分類し，それぞれをさらに 2 つの下位区分

に分け，下から A1，A2，B1，B2，C1，C2 という 6 段階に分類したものである。それぞれのレベルは，CAN-DO ディスクリプタ（能力記述文），つまり，言語を用いて何ができるかについての記述によって示されている。例えば，共通参照レベルの全体的な尺度として，A2 レベルの言語使用者は「ごく基本的な個人情報や家族情報，買い物，近所，仕事など，直接的関係がある領域に関する，よく使われる文や表現が理解できる。簡単で日常的な範囲なら，身近で日常の事柄についての情報交換に応ずることができる。自分の背景や身の回りの状況や，直接的な必要性のある領域の事柄を簡単な言葉で説明できる。」(吉島・大橋, 2004, p. 25) と記述されている。さらに，この 6 レベルについて，5 技能，つまりリーディング，リスニング，スピーキング（インタラクション，プロダクション），ライティングで実際に何ができるかという詳細な CAN-DO ディスクリプタが共通参照レベルの自己評価表として示されている。例えば，ライティングの A2 レベルには，「直接必要のある領域での事柄なら簡単に短いメモやメッセージを書くことができる」(吉島・大橋, 2004, p. 28) と記述されている。

(2) CEFR-J とは

CEFR-J とは，上記の CEFR に準拠しつつ日本の教育環境における英語教育に特化して開発されたものである。CEFR-J の特徴の一つは，レベルの細分化である。日本人英語学習者の多くが CEFR における A レベルであることから，特に下位レベルの細分化がおこなわれている。結果として，CEFR-J には，Pre-A1，A1.1，A1.2，A1.3，A2.1，A2.2，B1.1，B1.2，B2.1，B2.2，C1，C2 の合計 12 レベルが存在している。また，CEFR と同様に，5 技能の枠組みが存在する (投野編, 2013)。なお，CEFR-J において細分化されたレベルについても CAN-DO ディスクリプタが作成されている。例えば，聞くことの A2.1 の CEFR-J のディスクリプタは，「日常的・個人的な内容であれば，招待状，私的な手紙，メモ，メッセージなどを簡単な英語で書くことができる」である (投野編, 2013, p.294)。

1.3 CEFR と CEFR-J に基づく自己評価とパフォーマンス・テストに関する先行研究

ここでは，CEFR や CEFR-J に基づく自己評価やパフォーマンス・テストをおこなった調査や研究について説明する。平成 27 年度に文部科学省 (2016a, 2016b) によって英語教育改善のための英語力調査事業が実施され，中学校と高等学校で 4 技能のテストが実施された。その中で本研究が焦点を当てる書くことに関しては，中学校 3 年生も高等学校 3 年生も 2 種のライティングの課題が与えられた。技能テストの各得点域に対応する CEFR のレベルを決定するために，自己評価アンケートの結果を用いて閾値を設定した。具体的には，各得点域の生徒のうち，「英語 CAN-DO アンケート」における自己評価で該当レベルの項目に「できる」と回答した生徒の割合が中学校では 85%，高等学校では 60%を超えている点数を閾値として設定した。その結果，中学校 3 年生では A1 下位，A1 上位，A2 の 3 つのレベルで判定し， A1 上位，A2 の閾値はそれぞれ 35 点，70 点であった（95 点満点）。また高等学校 3 年生では A1，A2，B1，B2 の 4 つのレベルで判定し，A2，B1，B2 の閾値はそれぞれ 70 点，105 点，130 点であった（140 点満点）。そして，中学校 3 年生においては

997,154 名を対象にした結果，A2，A1 上位，A1 下位はそれぞれ約 0.2%（1,672 名），43.6%（334,664 名），56.2%（560,818 名）であった。なお，A1 下位には無回答者及び 0 点の生徒が含まれ，18.1%（14,303 名）がそれに該当した。また，高等学校 3 年生においては 81,494 名を対象にした結果，B2，B1，A2，A1 はそれぞれ約 0.0%（6 名），1.0%（15,081 名），18.5%（849 名），80.4%（65,558 名）であった。なお，A1 には無回答者及び 0 点の生徒が含まれ，12.5%（124,318 名）がそれに該当した。

文部科学省（2016a, 2016b）の評価方法には課題があると考えられる。この方法によると，レベルの閾値は自己評価によって判断されていることになる。しかし，CEFR や CEFR-J の CAN-DO ディスクリプタに基づく自己評価と外部試験の相関を検証した研究（伊庭, 2013; 渡慶次・Fewell・津嘉山・名城, 2016; 山下・阿佐美・新妻, 2016）によれば，自己評価と実際のパフォーマンスとの相関が弱いことが報告されているため，正しいレベル判定をおこなう上で，このように閾値の設定に自己評価を参考情報として活用してよいのかについては疑問が残る。

さらに，CAN-DO ディスクリプタへの自己評価と，実際に CAN-DO ディスクリプタを基に作成したタスクにおけるパフォーマンスとの関係を検証した研究がある。中谷（2013）では初級・中級レベルの大学生 65 名を対象とし，CEFR-J の CAN-DO ディスクリプタへの自己評価と，そのディスクリプタを基に設定されたタスクでの評価にどのような関係があるかを検証した。タスクは B1.1 と B1.2 の各ディスクリプタを基に設定した。また，タスクの評価については，ケンブリッジ英検の PET の評価基準と同様に 6 段階（0〜5 点）の全体的評価基準を用いてパフォーマンス評価をおこない，2 つのタスクの評価点の合計が 6 点以上を取った参加者を B1 レベルと判断した。結果として，該当レベルに達している人は CAN-DO ディスクリプタに対する自己評価と実際のパフォーマンスが一致するが，該当レベルに達していない人は，一致しない傾向があることが示された。さらにその原因として，中谷は「実際の目的のために書くというオーセンティックな体験が少ないことが影響を与えていると思われる。このため，未経験の上位のレベルのディスクリプタの内容の認識を，自分の書く能力と一致させるのは，それほど容易ではないと考えられる。」（pp. 270-271）と述べている。また，根岸（2013）は 38 名の中高生を対象にして「話すこと」の「やりとり」における自己評価と実際のパフォーマンスの関係を調べている。タスクは A2.1, A2.2, B1.1，B1.2 の各ディスクリプタを基に設定し，タスクの評価は 4 段階（1：全くできていない，2：あまりできていない，3：それなりにできている，4：問題なくできている）でおこなわれた。結果として，生徒の自己評価の結果と実際のパフォーマンスの一致率はそれほど高くなく，自己評価と実際のパフォーマンスの間には，関係は見出せなかった。また，各タスクの生徒の達成度を用いてタスクの困難度を算出したが，タスクの困難度の順は CEFR-J のレベル順と一致しなかった。これについては「同じテスト・スペックからテスト項目作成しても，異なった困難度になることはよくあること」（p. 218）とし，さらに「複数の変数を含むディスクリプタは，多様な解釈が可能であり，このために，そのディスクリプタに

対応するタスクが，困難度の観点からかなり異なる可能性があると考えられる」（p. 220）と述べている。すなわち，ディスクリプタを基にしてタスクを作成したしても，そのディスクリプタの解釈方法次第でそれに相応するタスクの難易度も変化してしまうことは起こり得るため，作成したタスクの困難度にも留意する必要がある。

1.4 階層性を想定した評価方法の提案

本研究では，中谷（2013）及び根岸（2013）の研究を参考にして，CEFR や CEFR-J のレベルに相当するライティング・タスクを 4 つ作成した。4 つのタスクにおけるパフォーマンスを評価することによって，当該学習者のレベルを特定するという方法をとった。

また，レベルの特定にあたっては，階層性を想定した評価方法を考案した。これは，文部科学省委託事業「中学校・高等学校における英語教育の抜本的改善のための指導方法等に関する実証研究（信州英語プロジェクト）」（信州大学, 2017; 酒井・阿部・菊原・木下・須野原, 2017）の研究で用いられた方法と同様のものである。CEFR や CEFR-J は，言語コミュニケーション能力を複数のレベルに尺度付けする枠組みであるが，上のレベルの言語行為をおこなうことができれば，下のレベルの言語行為をおこなえることを含意する。また逆に，下のレベルの言語行為ができなければ，その学習者は上のレベルの言語行為を遂行できないと推定される。つまり，CEFR や CEFR-J のレベルが階層性を示していると考えられる。この階層性に基づいて，次のように評価及びレベルの特定をおこなった。まず，本研究で作成した 4 つライティング・タスクの達成度を A，B，C の 3 段階で評価した。評価（判断）基準は，A は「該当レベルに達しており，さらに次のレベルに達している可能性がある」，B は「該当レベルに達している」，C は「該当レベルに達していない」である。3 段階にしたのは，学校教育において実施されている指導要録のための観点別学習状況の評価が 3 段階でおこなわれていることを参考にした。特に課題達成の可否を意識して採点をおこなうようにした。

すべてのタスクの達成度を評価した後，学習者のレベルの判定をおこなった。4 つのタスクの評価（3 段階）から，次の手順で学習者のレベルを推定した。例えば，B1.2 が A だったら，B2 以上という判定になる。B1.2 が B だったら，B1.2 という判定になる。さらに，B1.2 が C で，A2.1 が A ならば A2.2 という判定になる。

1.5 研究課題

本研究は，1.4 で説明したライティング・パフォーマンスの評価方法，すなわち CEFR や CEFR-J のレベルに相当するタスクを複数作成し，各タスクのパフォーマンスの 3 段階の評価に基づいて CEFR-J のレベルを特定する方法の妥当性を検証する。第 1 に，含意尺度の分析を用いて，CEFR や CEFR-J のレベルに対応するタスクにおける評価結果に階層性があるかどうか確認した。A2.1 が B であり，それより上のレベルのタスクの評価が C である場合，その学習者は A2.1 であると特定される。このレベルの特定が妥当であるか否かということを検討するために，A2.1 より下のレベルのタスクの評価に C が見られるという逆転現象がないかどうかを検証する。その検証方法として，2.3(3) で説明する含意尺度分析を

おこなう。

第 2 に，本研究の評価方法によるレベル判定と CAN-DO リスト形式の自己評価に基づくレベル判定との相関関係を検証した。自己評価との相関分析は，先行研究において，自己評価とパフォーマンス評価の相関が低いことが示されているが，この点を確認するためである。

第 3 に，ラッシュ分析により，4 つのタスクにおけるパフォーマンスの評価に基づいて算出した能力推定値と，本研究の評価方法によるレベル判定との相関分析をおこなった。前者は 4 つのタスクの 3 段階の評価情報をすべて用いて学習者の能力を推定する一方で，一般の学校においては活用しづらい。後者は C と評価されるタスクのレベルは何かに基づいて判断されており，使用する情報量に少ないが，方法が簡易であり，一般の学校で活用しやすいと考えられる。本研究の評価方法によるレベル判定が，どの程度，ラッシュ分析に基づく能力推定値を相関するのかを検討し，もし相関係数が高ければ，簡易である本研究の評価方法を用いることが勧められる。

まとめると，研究課題は以下の 3 つである。

1. 各タスクの評価において，どの程度階層性が確認できるか。
2. 本研究で提案する評価方法による判定は，CAN-DO リストに基づく自己評価の判定とどの程度相関するか。
3. 本研究で提案する評価方法による判定は，ラッシュ分析による能力推定値にはどの程度相関するか。

２．方法

2.1 参加者

本研究の参加者は，ある国立大学の共通教育科目の履修者である大学生である。受講生のうち，研究協力を承諾し，さらに欠席者及び欠損値のある者を除いた 80 名から得られたデータを分析対象とした。

2.2 材料

(1) ライティング・タスク

ライティング・タスクとして，Pre-A1，A1.2，A2.1，B1.2 の CAN-DO ディスクリプタを参考にし，4 つのタスクを設定した。

第 1 に，Pre-A1 相当のタスクとして，「英語の大文字・小文字をアルファベット順に書かせるタスク」を設定した。罫線が書かれた回答用紙の表に大文字を，裏に小文字を書かせた。時間は 10 分程度であった。

第 2 に，A1.2 相当のタスクとして，「自己紹介課題（自由）」を設定した。Tell your name. And make three sentences to introduce yourself. と指示した。時間は 5 分程度であった。

第 3 に，A2.1 相当のタスクとして，「自己紹介課題（トピック）」を設定した。自己紹介課題に，1 つのトピックについて書くように条件をつけたものである。トピックは参加者が

例えば「自分の住む町」や「好きなスポーツ」などを任意で選んだ。指示は，Choose one topic. Write your self-introduction about the topic.であった。時間は 5 分程度であった。

最後に，B1.2 相当のタスクとして，「論証文課題」を設定した。回答用紙の表に書かれた英語と日本語で指示を読み，回答用紙の裏の罫線上に英語を書くように指示をした。時間は 15 分程度であった。課題は TOEFL を参考に Uysal (2012) で用いられたものをより具体的な状況設定にしたものである。指示は次の通りである。

英語の指示

You are going to write the column for school newspaper for your sister school in America. The topic is “When people move to another country, they should adopt the customs and the lifestyles of the new country to succeed. Do you agree or disagree with the statement above?” Argue your position to convince an American reader by using strategies that you think are appropriate. Write about 100 to 150 words.

日本語の指示

あなたはアメリカの学校新聞の読者寄稿に投稿することになりました。テーマは「『他の国に引っ越すとき，そこで成功するためには自らを他の国の習慣や生活スタイルに適応させるべきだ』という主張があります。あなたはこの意見に賛成ですか，反対ですか」。あなたの意見を述べ，あなたが適切だと思う書き方で，アメリカ人の読み手を説得できるように書いてください。100 語から 150 文字程度で書いてください。

(2) 自己評価のための質問紙票

外的妥当性の検討のために，9 項目から成る自己評価の質問紙票を作成した（資料 A 参照)。A1.1，A1.2，A1.3，A2.1，A2.2 の 5 つのレベルの CAN-DO ディスクリプタについては，『中高生の英語学習に関する実態調査　2014』(ベネッセ教育総合研究所, 2014)を採用した[2]。また，Pre-A1，B1.1，B1.2，B2.1 の CAN-DO ディスクリプタについては投野編 (2013) から採用した。9 項目の CAN-DO リストについて，4 件法で回答させた（1: そう思わない，2: あまりそう思わない，3: ややそう思う，4: そう思う)。

2.3 採点方法と分析方法

(1) ライティング・タスクの採点方法

ライティング・パフォーマンスは，1.4 で述べた通りに採点をおこなった。A, B, C を判断する際の評価基準は，CEFR-J による各レベルの CAN-DO ディスクリプタを参考にして作成した。上述の通り，課題達成の可否を中心とし，文法の複雑性，語彙の幅などの観点から包括的な評価をおこなった。例えば，A2.1 レベルの課題「自己紹介課題（トピック)」であれば，1 つのトピックについて一貫して述べており，且つ and や but，because など基本的な接続詞を使用し，中学校レベルの文法を使いこなせている場合，「該当レベルに達している」，つまり B と判断した。また，B レベルの特徴に加え，高校レベルの文法の使用が認

められる場合や，表現の繰り返しが少なく語彙に多様性が見られる場合，「該当レベルに達しており，さらにその次のレベルに達している可能性がある」，つまり A と判断した。

なお，自己評価との相関を調べるために，判定されたレベルを数値化した。すなわち，Pre-A1 を 0 点，B2 以上を 8 点とし，各レベル 1 点増加で点数化をおこなった。

(2) 自己評価のための質問紙票の採点方法

自己評価の採点は，回答した番号に応じて「できる」もしくは「できない」という 2 つに分類した。つまり，自己評価において，3 もしくは 4 と答えた場合，参加者はそのレベルのタスクが「できる」と認識していると判断した。反対に 1 もしくは 2 と答えた場合，参加者はそのレベルのタスクを「できない」と認識していると判断した。自己評価における最終的なレベルの判定は，参加者が「できる」と認識している最も高いレベルをその参加者の自己評価におけるレベルとした。例えば，A2.1 の CAN-DO ディスクリプタに，3 と答え，A2.2 以上のレベルで 2 や 1 と答えた場合は，参加者は自己評価において A2.1 レベルであると判定した。また，B1.1 で 4 と回答し，A2.1 で 2 と回答している場合は，より高いレベルの方を採用し，B1.1 と判定した。

パフォーマンス評価との相関を調べるために，判定されたレベルを数値化した。すなわち，Pre-A1 を 0 点とし，B2 以上を 8 点とし，各レベル 1 点増加で点数化をおこなった。

(3) 分析

研究課題 1 に関して，含意尺度法（Hatch & Lazaraton, 1991）を用いた階層性の検討をおこなった。Hatch and Lazaraton (1991) によれば，階層性の判定には，coefficient of reproducibility, coefficient of minimum marginal reproducibility, percent improvement, coefficient of scalability の 4 つの数値を用いる。coefficient of reproducibility は 0.90 より大きいこと，coefficient of minimum marginal reproducibility が 0.945 以下であること，そして coefficient of scalability（含意尺度係数）が 0.60 以上で階層性が認められるとしている。

研究課題 2 については，ライティング・パフォーマンス評価の判定と自己評価の判定の相関分析をおこなった。2.3 の手順で算出した得点を順位相関係数である Kendall の τ を用いて相関分析をおこなった。

研究課題 3 については，パフォーマンス評価と自己評価によるレベル判定を基に算出した数値と，ラッシュ分析による実際のパフォーマンスと自己評価による能力推定値を順位相関係数である Kendall の τ 及びピアソンの関立相関係数を用いて相関分析をおこなった。なお，ラッシュ分析においては，WinSteps (3.91.2) の部分クレジットモデル（partial credit model）を用い，実際のパフォーマンス評価の分析では，A，B，C をそれぞれ 0，1，2 と変換し，自己評価の分析には回答の 4 件法（1，2，3，4）の値を用いて分析をおこなった。

３．結果と考察

3.1　パフォーマンス評価における階層性について

まず，ライティング・タスクにおけるパフォーマンスを評価した結果を表 1 に示す。A 評価を見てみると，Pre-A1 の 72 名から徐々に減っており，B1.2 では 5 名であった。B 評価と C 評価の分布は，タスクのレベルによって異なるが，B1.2 が最も多く，44 名であった。

なお，ライティング・タスク評価基準による 2 人の採点者による 3 段階評価の一致率は，Pre-A1 は 100%，A1.2 は 87.7%，A2.1 は 86.4%，B2.1 は 77.8%であった。不一致であった評価については，2 人の評価者が議論をおこない，最終的な評価を決定した。一致度は，タスクの難易度が上がるにつれて低くなったことから，タスクのレベルが高くなる時には，採点者間信頼性を確保するための工夫が必要であると考えられる。

表 1　各タスクにおける評価の分布

	A 評価	B 評価	C 評価
Pre-A1	72	0	8
A1.2	26	54	0
A2.1	15	58	7
B1.2	5	31	44

次に，4 つのライティング・タスクにおけるパフォーマンスの評価の階層性を検討するために行われた含意尺度分析を行った（表 2 参照）。coefficient of reproducibility は 0.950, coefficient of minimum marginal reproducibility は 0.859，coefficient of scalability（含意尺度係数） は 0.644 となり，パフォーマンス評価における階層性が確認された。これにより，研究課題 1 に関して，上のレベルのタスクにおけるパフォーマンスの評価が B 以上であった場合，下のレベルのタスクにおけるパフォーマンスの評価が B 以上である可能性が高く，また逆に，下のレベルのタスクにおけるパフォーマンスの評価が C であった場合，上のレベルのタスクにおけるパフォーマンスの評価が C である可能性が高いことが示された。すなわち，4 つのタスクにおけるパフォーマンス評価は階層性を示しており，その階層性を想定してレベルを判定する方法は妥当であるといえる。

表 2　含意尺度係数

指標	基準	結果
Crep	> .900	0.950
MMrep	≦ .945	0.859
% improvement		0.091
Cscal	> .600	0.644

3.2　パフォーマンス評価と自己評価の判定の相関について

表 3 は，ライティング・タスクのパフォーマンスや自己評価に基づいて判定したレベル

の分布をまとめている（判定方法については 2.3 を参照）。ライティング・タスクのパフォーマンスの判定では，A2.1（39 名）が最多となり，続いて B1.2（28 名），B2.1 及び A1.2 及び A2.2/B1.1（4 名），A1.3（1 名）であった。一方，自己評価による判定では，B1.1（26 名）が最多となり，続いて A1.3（15 名），B1.2（14 名），A2.2（13 名），B2.1（9 名），A1.2（2 名），A1.1（1 名）であった。

表 3 から，パフォーマンスの評価判定と自己評価の評価判定にずれが見られることかわかる。例えば，自己評価で A1.1 と判定された参加者 1 名は，パフォーマンスの評価判定では B1.2 と判定されている。

ライティング・タスクの評価基準によるレベル判定の得点と自己評価によるレベル判定の得点の順位相関は，.229（p = .016）であった。これにより，研究課題 2 に関し，統計的に有意であったものの，相関係数は小さかったことが確認された。これは，先行研究で示された相関係数の低さを確認するものとなった。ただし，自己評価に用いられた項目の一部はベネッセ教育総合研究所（2014）の Can-do statements の記述であり，これらの示す自己評価のレベルが，CEFR や CEFR-J のレベルと正確に対応しているか否かは議論の余地がある。つまり，自己評価に用いられた記述がベネッセ教育総合研究所（2014）のものであったことが，パフォーマンス評価と自己評価の判定の相関係数を小さくした可能性がある。

表 3　自己評価及び評価基準によるレベルの判定と人数（人）

自己評価＼パフォーマンス	A1.2	A1.3	A2.1	A2.2/B1.1	B1.2	B2.1	総計
A1.1					1		1
A1.2			2				2
A1.3	2	1	9		3		15
A2.2			7		6		13
B1.1	1		12	3	9	1	26
B1.2	1		7		4	2	14
B2.1			2	1	5	1	9
総計	4	1	39	4	28	4	80

3.3　パフォーマンス評価と自己評価のラッシュ分析について

タスクにおけるパフォーマンスの評価を，ラッシュ分析を用いて間隔尺度に変換した結果が図 1 にまとめられている。図の右側は，本研究において作成された各レベルのタスクが項目困難度のロジット値に従って配置されている。ロジット値が高ければ，困難度が高いことを示している。図 1 によれば，B1.2 のタスクが最も困難であり，A2.1，A1.2，Pre-A1 の順に困難度が下がっていることが示されている。したがって，本研究において作成されたタスクの難易度は CRFR-J の想定と同じ並び方を示した。また，図の左側は，参加者が能

力推定値のロジット値に従って配置されており，「#」のマークは 3 人，「.」のマークは 1 人または 2 人の人数を示している。ロジット値が高ければ，能力が高いことを示している。本研究の参加者 80 名のうち 36 名（45.0%）が B1.2 と A2.1 の間に位置しており，27 名（33.8%）が A2.1 と A1.2 の間に位置していることが示された。B1.2 以上に位置する参加者は 9 名（11.3%）であった。

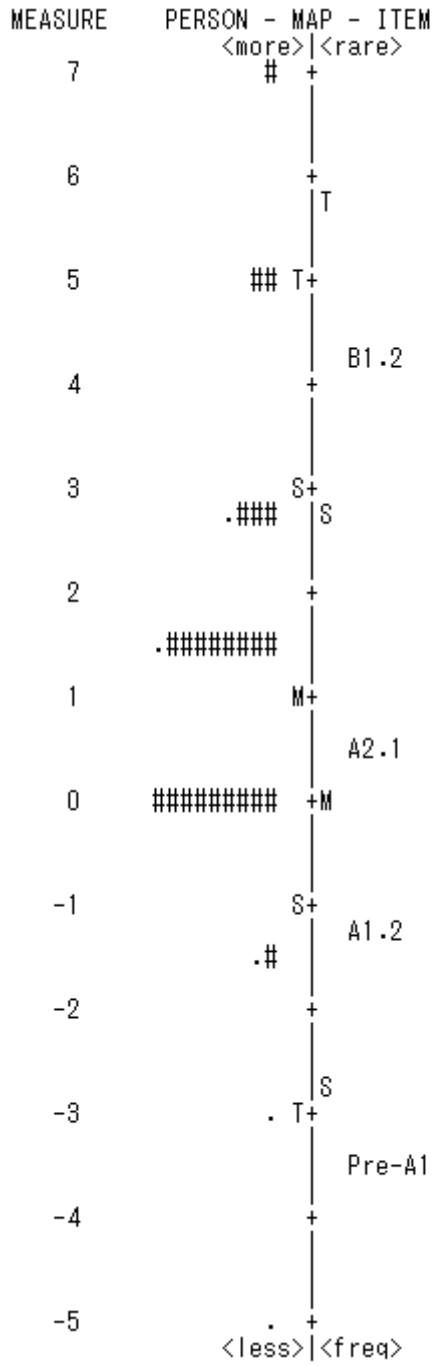

図 1　パフォーマンス評価のラッシュ分析

図 2 は，ラッシュ分析を用いて，自己評価の回答（順位尺度）を間隔尺度（項目困難度や能力推定値）に変換した結果を示している。図の右側は，自己評価における各レベルの質問項目が項目困難度のロジット値に従って配置されている。図 2 によれば，B2.1 が最も困難であることが示されている。また，A1.1 と Pre-A1 が最も簡単なタスクであり，同程度の困難度であることが示されている。順番を見ると，A2.1 の困難度が，B1.1 の困難度よりも高く，CEFR-J の想定と異なっている。図の左側は，参加者が能力推定値のロジット値に従って配置されており，図中の「X」のマークは 1 人の人数を示している。ロジット値が高ければ，自己評価の判断に基づく能力が高いことを示している。

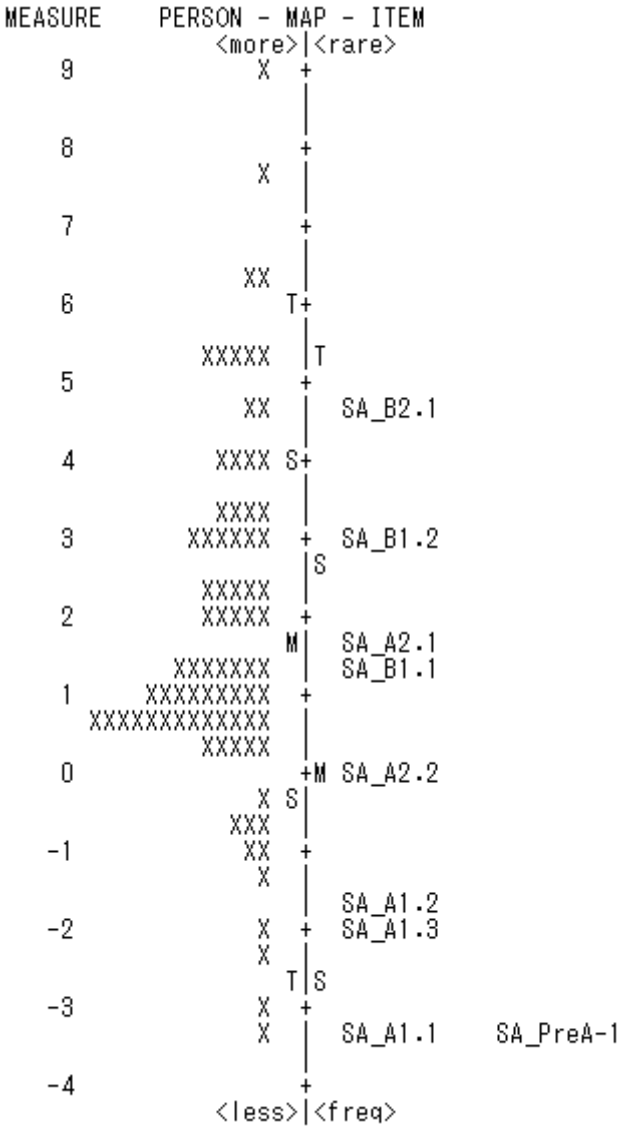

図 2　自己評価のラッシュ分析

A2.1 のレベルが B1.1 よりも難しいと判断される傾向にあった理由として，自己評価の CAN-DO ディスクリプタが参加者にとって身近なものではなかったためと考えられる。A2.1 の CAN-DO ディスクリプタが「英語の手紙や電子メールなどで，ある程度まとまった内容を，それほど辞書を引かなくても書くことができる。」であり，B1.1 の CAN-DO ディスクリプタは「身近な状況で使われる語彙文法を用いれば，筋道を立てて，作業の手順などを示す説明文を英語で書くことができる」であった。A2.1 に書かれている「英語の手紙や電子メールなど」は，日本人学習者が普段おこなわないため，自己評価で否定的な回答をした可能性が考えられる。このことは，根岸（2006）が「経験に基づかない自己申告による Can-do Statements には，高い信頼性は望めないだろう」（p. 101）との指摘に一致する。

研究課題 3 に関して，本研究が提案するレベル判定と，ラッシュ分析の結果の相関分析を行った。2.3 で述べた判定方法の特徴は，あるレベルについて，1 つのタスクのパフォーマンスのみで判定しようとするものである。一方，ラッシュ分析においては，すべてのタスクや項目における回答を基にして，能力推定値を計算する。ラッシュ分析の結果と比較することによって，この判定方法について考察する。

パフォーマンス評価のレベル判定とパフォーマンス結果に基づく能力推定値の相関係数（Kendall の τ）は，.774（p = .000）であった。すなわち，比較的高い相関係数が得られ，統計的に有意であった。このことから，1 つのタスクのパフォーマンスのみでレベル判定をおこなうことは，すべてのタスクのパフォーマンス結果から推定する能力推定値と統計的に有意で高い相関係数が得られたことから，ラッシュ分析でタスクの難易度の検証に問題がない場合，妥当であると考えられる。

自己評価のレベル判定と自己評価の回答に基づく能力推定値の相関係数（Kendall の τ）は，.687（p = .000)であった。自己評価についても，比較的高い相関係数が得られ，また，統計的に有意であった。今回，3 以上の回答を示したか否かによって自己評価のレベルの判定をおこなったが，これは，すべての項目の 4 件法の回答に基づいて推定する能力推定値と，統計的に有意で高い相関係数が得られたことから，自己評価のレベル判定も妥当であったと考えられる。

3.2 で示したように，ライティング・タスクの評価基準によるレベル判定の得点と自己評価によるレベル判定の得点の順位相関は，.229（p = .016）であり，低かった。そこで，ライティング・タスクの評価のレベル判定と，自己評価の回答に対してラッシュ分析を通して得られた能力推定値の間の相関係数（ピアソンの相関係数）を計算した。その結果，相関係数は r = .460（p = .000）であり，統計的に有意であり，中程度の相関が得られた。1 つのタスクのパフォーマンスのみのレベル判定と 3 以上の回答を示したか否かによる自己評価のレベル判定の相関係数に比べて，パフォーマンス評価のレベル判定と自己評価の能力推定値との相関係数は高くなった。これは，パフォーマンス評価及び自己評価の際に，「できる」「できない」のみで判断してレベルを判定せずに，それぞれ学習者の能力を 3 段階（0，1，

2）4 段階（0，1，2，3）で判断したことに起因する可能性がある。

４．結論

本研究では，CEFR 及び CEFR-J のレベルに階層性があると想定されることに基づいて提案された評価方法の妥当性について，含意尺度を用いて階層性の検証をおこなった。また，本研究の評価方法によるレベル判定と CAN-DO リスト形式の自己評価との相関を検証した。本研究の評価方法による判定とラッシュ分析による能力推定値にはどのような関係があるかについても検証をおこなった。

その結果，本研究の評価方法による判定において，含意尺度係数の観点から階層性が確認された。また，本研究の評価方法によるレベル判定と CAN-DO リスト形式の自己評価との相関係数は統計的に有意に低いことが確認された。さらに，本研究の評価方法による判定とラッシュ分析による実際のパフォーマンスの能力推定値の相関は比較的高いことが確認された。同様に，自己評価とラッシュ分析による自己評価の能力推定値の相関も比較的高いことが示された。さらに，ラッシュ分析による実際のパフォーマンスと自己評価の能力推定値の間に中程度の相関があることが示された。

したがって，本研究で提案した評価方法について，CEFR や CEFR-J のレベルを参考とした複数のタスクを作成し，A，B，C の 3 段階で評価し，レベルを判定する方法には，ラッシュ分析でタスクの難易度の検証に問題がない場合，妥当性があると言える。また，自己評価については，今回用いた CAN-DO リストの一部はベネッセ教育総合研究所（2014）の項目であり，CEFR-J の項目だけではなかったことが自己評価の正確さを下げた一因になっていることは否定できないが，先行研究で示されている通り，学習者は自分のレベルを正確に認識することができない傾向にあると言える。

本研究のライティング課題は，Pre-A1，A1.2，A2.1，B1.2 の 4 レベルであり，A1.3 と B1.2 の間のみレベルが 2 つ空いている。信州大学 (2017) 及び酒井他 (2017) では，CEFR-J の A1.1，A1.3，A2.2，B1.2 レベルに相当するスピーキング・タスクを作成し評価をおこなっている。これは，タスクのレベルが一つ飛ばしになっているため，その間のレベルも想定し，9 レベルを用いた含意尺度による分析をおこなうことができる。しかしながら，本研究では，A2.1 の B1.2 の間に 2 つのレベルが存在しているため，4 レベルのみを用いた含意尺度分析をおこなった。ライティング・パフォーマンス評価においても，作成したタスク以外のレベルを推定し階層性の分析をおこなうためには，適切なレベルのタスクの設定が求められるだろう。

また，本研究の Pre-A1，A1.2，A2.1 のタスクについて，C と判定された参加者はそれぞれ，8 名, 0 名, 7 名であった。本研究の参加者は大学生であったため，レベルの比較的低いタスクについてはほとんどの参加者が達成できていた。今後は，参加者のレベルに合わせたより適切なレベルのタスクを作成する必要があるだろう。

本研究では階層性が想定される CEFR や CEFR-J のレベルを参考としてタスクを設定し

た。これらの指標と同様に階層性が想定される他の指標を用いての追研究も，本研究で検証した評価方法を他の指標で運用する際には，その妥当性を確認するために必要であろう。

注

[1] 本研究は，第 47 回中部地区英語教育学会（平成 29 年 6 月 24 日於信州大学教育学部）で口頭発表した「ライティング・パフォーマンス評価方法の検討―含意尺度の分析と CAN-DO リストの自己評価との相関分析―」を基に執筆したものである。

[2] ベネッセ教育総合研究所 (2014) の CAN-DO リストを採用した主たる理由は，本研究の計画段階では『中高生の英語学習に関する実態調査　2014』の結果と本研究の結果を比較することを予定していたからである。なお，ベネッセ教育総合研究所 (2014) の項目は，ベネッセホールディングス（2011）の項目を用いている。ベネッセ教育総合研究所 (2014) の項目は A1.1 から A2.2 までの 5 項目しかないため，項目を補充するために投野編 (2013) を参考にした。

謝辞

本研究は，基盤研究(C) 15K02785「英語コミュニケーション能力のパフォーマンス評価法の理論的・実証的研究」（研究代表者酒井英樹）の助成を受けている。

引用文献

ベネッセホールディングス（2011）. *GTEC for Students can-do statements.* Retrieved from http://www.benesse-gtec.com/fs/pdf/ab_feedback/can-do_statements.pdf

ベネッセ教育総合研究所（2014）.『中高生の英語学習に関する実態調査　2014』Retrieved from http://berd.benesse.jp/global/research/detail1.php?id=4356

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Retrieved from https://www.coe.int/t/dg4/linguistic/source/framework_en.pdf（吉島茂・大橋理枝訳（訳・編）(2004).『外国語教育Ⅱ―外国語の学習，教授，評価のためのヨーロッパ参照枠―』東京: 朝日出版社）

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Bostan, MA: Heinle & Heinle Publishers.

伊庭緑 (2013).「ヨーロッパ言語共通参照枠 (CEFR) に基づく学生調査の結果と英語力試験の相関性：学生評価を大学の質保証につなげる試みの中で」『言語と文化』第 17 号, 61-80.

文部科学省（2016a）.『平成 27 年度 英語教育改善のための英語力調査事業（中学校）報告書』Retrieved from

http://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2016/12/16/1375533_5.pdf

文部科学省 (2016b). 『平成 27 年度 英語教育改善のための英語力調査事業（高等学校）報告書』Retrieved from http://www.mext.go.jp/component/a_menu/education/detail/_icsFiles/afieldfile/2016/12/16/1375533_1.pdf

中谷安男 (2013). 「Q 41　Writing の CAN-DO と 実際のスキルとの関連性は?」 投野由紀夫（編） 『英語到達度指標 CEFR-J ガイドブック』(pp. 265-272). 東京: 大修館書店.

根岸雅史 (2006). GTEC for STUDENTS Can-do Statements の妥当性検証研究概観. *ARCLE Review, 1,* 96-103.

根岸雅史 (2013). 「Q 35　Spoken Interaction の CAN-DO と 実際のスキルとの関連性は?」 投野由紀夫（編） 『英語到達度指標 CEFR-J ガイドブック』(pp. 212-220). 東京: 大修館書店.

酒井英樹・阿部敏子・菊原健吾・木下愛里・須野原美香 (2017). 「含意尺度分析及び自己評価と外部試験との相関分析に基づくスピーキング・パフォーマンス評価方法の検討」『第 43 回全国英語教育学会島根研究大会発表予稿集』, 142-143.

信州大学 (2017).「中学校・高等学校における英語教育の抜本的改善のための指導方法等に関する実証研究（信州英語プロジェクト）」Retrieved from http://shinshu-eigoproject.jp/about

竹内理・水本篤 (2012). 『外国語教育研究ハンドブック』東京: 松柏社.

渡慶次正則・Fewell, N.・津嘉山淳子・名城義久 (2016). 「CEFR-J に基づいた Can-Do ディスクリプタの信頼性と相関関係の基礎的調査：M 大学教養英語の事例」『名桜大学総合研究』第 25 号, 13-23.

投野由紀夫（編）(2013).『英語到達度指標 CEFR-J ガイドブック』東京: 大修館書店.

Uysal, H. H. (2008). Tracing the culture behind writing: Rhetorical patterns and bidirectional transfer in L1 and L2 essays of Turkish writers in relation to educational context. *Journal of Second Language Writing*, *17*, 183–207.

山下早代子・阿佐美敦子・新妻奈緒美 (2016). 「英語運用能力に対するセルフ・アセスメント(自己評価)と TOEIC スコア」『実践女子大学人間社会学部紀要』第 12 集, 81-96.

資料 A. 自己評価の CAN-DO リスト

Pre-A1	アルファベットの大文字・小文字，単語のつづりをブロック体で書くことができる。
A1.1	自分の名前や住んでいる場所などの内容を含む，簡単な自己紹介文を英語で書くことができる。
A1.2	絵はがきやカードに簡単な英語のメッセージを書くことができる。
A1.3	英語の手紙や電子メールなどを，辞書を引きながらであれば，書くことができる。
A2.1	英語の手紙や電子メールなどで，ある程度まとまった内容を，それほど辞書を引かなくても書くことができる。
A2.2	自分の興味のある話題やものに関して，英語で意見や感想を短く書くことができる。
B1.1	身近な状況で使われる語彙・文法を用いれば，筋道を立てて，作業の手順などを示す説明文を英語で書くことができる。
B1.2	新聞記事や映画などについて，専門的でない語彙や複雑でない文法構造を用いて，自分の意見を含めて英語であらすじをまとめることができる。
B2.1	自分の専門分野であればメールやファックス，ビジネス・レターなどの英語のビジネス文書を，自分の意図を含めながら，適切な文体で書くことができる。【修正あり】

（2017年1/月 /日　受付）
（2018年 2月 4日　受理）