EIKEN BULLETIN

vol.30 2018

第30回「英検」研究助成　報告

A. 研究部門
- 英語能力テストに関する研究

B. 実践部門
- 英語能力向上をめざす教育実践

C. 調査部門
- 英語教育関連の調査・アンケートの実施と分析

公益財団法人 日本英語検定協会

第30回「英検」研究助成　選考委員

（役職は委嘱当時，＊印は専門選考委員）　※五十音順

＊池田　　央　立教大学名誉教授・(株)教育測定研究所 取締役

石鍋　　浩　全日本中学校長会 会長代理・総務部長

＊伊藤　寛和　公益財団法人 日本英語検定協会 理事・事務局長

＊大友　賢二　筑波大学名誉教授

＊長　　勝彦　一般財団法人 語学教育研究所 参与

＊小池　生夫　慶應義塾大学・明海大学名誉教授

栃倉　和則　全国英語教育研究団体連合会 会長

水野　次郎　全国高等学校長協会 英語科・国際科部会 理事長

＊村木　英治　東北大学名誉教授・(株)教育測定研究所 取締役

＊吉田　研作　上智大学特別招聘教授・言語教育研究センター長

＊和田　　稔　明海大学名誉教授

読者を考えた研究を

～ 第30回「英検」研究助成 報告に寄せて ～

上智大学特別招聘教授・言語教育研究センター長

吉田 研作

毎年「英検」研究助成の申請書を読ませていただいているが，応募される大学院生や教員の熱意が伝わってくるものが多く，審査をする側にも考えさせられる計画書がいくつもある。ただ，その中で少し気になる点があるので，それについて述べることとする。

応募される研究計画は，研究者（主に大学院生）による英語の修得や評価に関する基礎的なものから日常の授業や教育方法に関する具体的なものまで，非常に幅が広い。ただ，研究者が提出する研究計画は立派なものが多い反面，一般の英語教師，あるいは，読者にとって果たしてどこまで理解できるだろうか，と疑問を持たざるを得ないものが多々ある。最終的には実践に結びつけることを念頭に置いたものがほとんどだが，その前の研究方法や結果分析，結果のディスカッションの段階で，読むのを諦めてしまう読者がいるのではないかと心配する。

その一方で，英語教育の実践や調査を前面に出した具体的な計画書で現場の教師が中心になって考えられたものがある。読者にとっては分かりやすいものが多いのは確かだが，その反面，あまりにも個別の状況に依拠しているために，状況が変わったら適応できないのではないか，と思われるものがある。つまり，研究の汎用性について疑問を持ってしまうのである。それぞれの現場で具体的で個別の課題を抱えておられることは良くわかるが，それが，自分たちだけでなく，より広範な状況においても使えなければ，校内研究に留まってしまう可能性があるのである。

「英検」研究助成の特徴は，学問研究，調査，実践等，色々な分野の研究を取り入れているところにある。しかし，共通していることは，どの分野の研究も，様々な背景を持った読者に理解できるものであり，また，その結果が様々な教育場面に適応できることである。

上記の点を十分考慮して計画を立てていただきたいと思う。

Contents

● **はじめに**

～ 第30回「英検」研究助成 報告に寄せて ～

上智大学特別招聘教授・言語教育研究センター長　吉田 研作 …… 3

● **報告別講評**　専門選考委員

吉田 研作／村木 英治／大友 賢二／池田 央／長 勝彦／小池 生夫／和田 稔 …… 6

A 研究部門

英文読解におけるモニタリング能力の測定
― 一貫性と結束性の比較を通して ―

研究者：茨城県／筑波大学大学院　在籍　小木曽 智子 …… 12

多肢選択の解答パターンに基づく未知語推測プロセス診断:
形態素・文脈情報の活用から

研究者：茨城県／筑波大学大学院　在籍・日本学術振興会特別研究員　神村 幸蔵 …… 27

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 …… 42

スピーキング練習における特定文法項目に焦点を当てた
直接的修正フィードバックの効果検証

研究者：愛知県／名古屋大学大学院　在籍　小林 真実 …… 67

パラフレージング・レベルによる錯乱肢の作成
― 誤答分析に基づくテキスト理解度の診断に向けて ―

研究者：茨城県／筑波大学大学院　在籍　政所 里佳 …… 85

B 実践部門

CAN-DOリストとCAN-DO Check Sheetの導入が高校生英語学習者の自己調整学習と英語力向上に及ぼす効果

研究者：兵庫県／兵庫県立川西緑台高等学校　教諭　津田 敦子 …… 140
申請時：神戸大学附属中等教育学校　教諭

ルーブリックを用いて段階を踏んだ，継続的なスピーキング活動の実践

研究者：埼玉県／川口市立高等学校　教諭　鶴田 京子 …… 161
申請時：川口市立県陽高等学校　教諭

高校生を対象にした「書くこと」に焦点をあてた協働学習が英文エッセイにおけるCEFRレベル別語彙使用に与える効果

研究者：愛媛県／愛媛県立西条農業高等学校　教諭　三好 徹明 …… 177
申請時：愛媛大学附属高等学校　教諭

C 調査部門

LFCにおけるcore項目の学習:日本人英語学習者に関する音響分析

研究者：神奈川県／慶應義塾高等学校　教諭　北川　彩 …… 204

発話の諸側面に対する意識の質問紙尺度の開発と妥当性の検証:発話生成モデルの観点から

研究者：英国／ランカスター大学大学院 博士課程　在籍　鈴木 駿吾 …… 221

報告別講評

評　者	A	B	C
吉田 研作	研究部門・報告Ⅰ		調査部門・報告Ⅰ
村木 英治	研究部門・報告Ⅱ 研究部門・報告Ⅴ		
大友 賢二	研究部門・報告Ⅲ		
池田　央	研究部門・報告Ⅳ	実践部門・報告Ⅳ	
長　勝彦		実践部門・報告Ⅰ	
小池 生夫		実践部門・報告Ⅱ	調査部門・報告Ⅱ
和田　稔		実践部門・報告Ⅲ	

A. 研究部門・報告Ⅰ

吉田 研作

英文読解における モニタリング能力の測定

―一貫性と結束性の比較を通して―

【研究者：小木曽 智子】

本研究は，日本人英語学習者の英文読解の際の読解ストラテジーの何がどのように使われているかを見たものである。従来，読解においては，テキストの難易度を測定するために，いわゆるReadabilty scaleが使われてきた。しかし，本研究では，読解に必須とされる文章の結束性および意味的一貫性という基準を加えることにより，より精度の高い読解力指標の策定を試みている。

研究の結果，学習者のレベルに合わせた教材選定を行うに当たり，結束性の尺度を取り入れることが有効である可能性が示唆された。また，学習者の読解中のプロセスを見ると，結束性に対するモニタリングはあまり行っていないが，意味的な一貫性についてはモニタリングを行っていることが分かった。この一つの解釈としては，学習者は読んでいる文章の結束性に多少の問題があってもそれに気づかず，意味的な一貫性を見出そうとする心理が働いていることが考えられる。この結果は，今後の読解指導において大切な意味を持っているように思う。つまり，結束性が分からなければ論理的に構成された文章を読む際には弊害となる可能性がでてくるのであり，今後の読解指導で強調しなければならないのである。

A. 研究部門・報告Ⅱ

村木 英治

多肢選択の解答パターンに 基づく未知語推測プロセス診断： 形態素・文脈情報の活用から

【研究者：神村 幸蔵】

我々が新しい英語文を読解する際，未知の英単

語に遭遇する場合がある。そんなとき辞書を引いてその意味を探るのが常道の作業だ。しかしその辞書が手元にない場合がある。あっても英語の単語数はその長い歴史性とその広い国際性で膨大なものとなっている。Oxford Dictionaryはどこにでも手軽に持ってはいけない。それに英語読解の学力を測る試験には辞書持ち込み可能の例は少ない。「未知の英単語に遭遇する場合」は決して英語の初学者だけの経験ではない。むしろ英語力が進むにつれ専門性が高い文章を読む機会が生じ、未知語に遭遇することが格段に増え，それら未知単語の意味を推測しなければならない必要性が増大する。神村氏が対象として選んだ英語学習者の未知語推測プロセスは英語学習プロセスの研究の重要なトピックである。

神村氏は多肢選択質問項目を用いた英文読解テストにフォーカスし，その項目反応に使われる未知語推測プロセスにおいて，学習者が用いる未知語である目標語の形態素についての情報、特に接尾語などが含まれる拘束形態素とそれが使われている文脈からの情報のふたつに注目する。しかしこのプロセスが根本的であるが故に研究には力技と努力が必要だった。この絡み合った二つの情報源をどう切り離して研究していくのか。また被験者それぞれの形態素分析力の差や単語力や読解力の個人差も考慮しなければならない。私は神村氏の過去研究のReviewに多くのことを教えられ，また一つ一つ問題の曖昧さをつぶしていくような綿密な研究の準備にも感心した。

研究の3つの結論のひとつとして神村氏は「形態素情報に依存した推測をしがちな学習者は，自身の英語熟達度の低さ故に文脈情報を利用することが難しく，推測の成功にたどり着くことができない」と述べた。英語学習者のすべてのレベルで行っているこの未知語推測基本プロセスの奥の深さと面白さを気づかさせてくれた秀逸の論文である。

A. 研究部門・報告Ⅲ

大友 賢二

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：久保田 恵佑】

この研究のタイトルに関しては，その理解を頂いた方が，ことの進み具合としては良いと考える。まず，RTW，EBBという用語である。RTWというのは，Reading-to-Writing Task の中のRTWを意味している。EBB というのはEmpirically derived，Binary-choice，Boundary-definition scale のEBBを示すもので，テスト採点手順の方法を示すものである。しかし，なぜ，このような厄介な手順で書く能力を測定する方向が増えてきているのであろうか。そのことは，Task-based Language Teaching など，「現実の世界で使えるtask に視点」を置く現実の指導法に注目している状況を見ればよい。そこではこの厄介な手順を踏んだがlistening やspeaking，つまり，「技能統合的タスク」の能力測定が必要であり，なぜ，このような厄介な能力測定法が登場してきたのか，その意味が理解できるはずである。こうした教授法の現状を捉えた論題を取り上げている久保田氏に対し大きな拍手を送りたい。

このような，RTWタスクなどを測定する場合のルーブリックに関しては，どんなものが適切であるかを求めようとしているのが，本論の狙いである。具体的な研究課題は3つある。そして，その研究結論の一つは，（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と言うことなどがある。

この研究に対する課題や提案として，まず，第1に挙げられるのは，「サンプルサイズ」の課題である。ここでは，「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。」「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」と述べている。「サンプルサイズをいくつにしたら良いか」と質問を受けることが多い。多角的検討がぜひ必要であろ

う。課題の2は，ここでは、技術統合的ライティングを用いているが，R=>Wの作業を受験者が行いWの能力を測定している。しかし，その場合のWは、L=>W のWの能力とは違うであろう。あるいは同じなのか。こうした場合のWと言う能力の分類と統合をどう考えたら良いであろうか。第3の課題は，ルーブリックの評価観点の検討である。TEAPでは，ライティングの評価観点は，main ideas, coherence, cohesion, lexical range and accuracy, grammatical range and accuracy（和泉(2016:255)）としてある。これに対応するEBBではTEAPでの coherence と cohesion とを1本化している点が違う。さらに，TEAPルーブリックとEBBルーブリックの評点観点と「レベル数」の違いが，診断的評価に適しているか否かの決定的要因となるのであろうか。こうした課題のさらなる究明が必要であろう。第4の課題は，ここでは，EBB scaleがTEAP scale より優れているとの結果が見られたが，EBB scale が，(Upshur & Turner, 1995) で提案されて10年過ぎた今日でもあまり使用されていない。その理由は，(1) empirical investigation of EBB scale, (2) developing an EBB scale is laborious work ではないかという声(Hirai & Koizumi, 2008: 15) があることも忘れてはならない。

A. 研究部門・報告Ⅳ

池田　央

スピーキング練習における特定文法項目に焦点を当てた直接的修正フィードバックの効果検証

【研究者：小林 真実】

これは英語のスピーキング指導で学習者の誤りを指摘し修正を促す指導で3つの方法を比較し，どれがよいか検討する研究である。ここでは特に動詞の過去形に焦点を当てている。

その場合どこが間違えているか。過去形なのか，定動詞の欠落なのか目的語や前置詞なのか等々，焦点を当てて修正フィードバックを行なわせるのと，そうした焦点を与えないで修正させるのと，また，そうした修正の時間的感覚を変えたグループとの違いなどを比較したりする興味ある研究である。

しかし結果としては，あまり群間で有意差が見られなかったケースが多かったようである。それは思うに比較するグループの人数が15〜17人と少なかったからではないかと思われる。今後の研究が待たれる問題であろう。

A. 研究部門・報告Ⅴ

村木 英治

パラフレージング・レベルによる錯乱肢の作成

— 誤答分析に基づくテキスト理解度の診断に向けて —

【研究者：政所 里佳】

私の専門である項目反応モデルでもすべての能力レベルにおいて一様のguessing parameterを含む3PL model離れが進んでいるという。確かに多肢選択項目に各錯乱肢の選択確率は学力レベルとのinteractionと見なすほうが適切であるし，このように考えることはPsychometricianが目指している錯乱肢を含めての多肢選択項目のAI応用のcomputer generationにおおいに益するに違いない。パラフレージング・レベルを変数として作成した錯乱肢と正答肢選択確率やその魅力度の関係を研究のテーマとして選んだ政所氏はテスト学の未来を確かに見据えている。

政所氏の本論文はふたつの調査から成り立っている。最初の調査で政所氏はTOEICやTOEFLなどのいくつかの現行の読解テストの錯乱肢や正答肢をパラフレーズ・レベルの観点から分析している。その結論を踏まえ調査2ではパラフレーズ・レベルを意図的に変えた独自の読解テストを作成し，受験者の英語能力のレベルと各選択肢の質的かつ具体的で詳細な関係を明らかにしていく。特に政所氏の4つの主なる結論と同じようなトピックを扱った過去論文との比較検証は大変面白く示唆されることが大きい。

政所氏は最後に選択肢に対する回答自信度と正誤判断を組み合わせた魅力度を本研究で導入した意義を強調し、今後の課題としてその方面の研究

の可能性をおおいに示唆した。Psychometricianとしてテストのこのような心理学的アプローチの将来性に私もおおいに期待している。

B. 実践部門・報告Ⅱ

小池 生夫

CAN-DOリストとCAN-DO Check Sheetの導入が高校生英語学習者の自己調整学習と英語力向上に及ぼす効果

【研究者：津田 敦子】

この論文は，自分たちでCAN-DOリストやCAN-DO Check Sheetを作成し，それを授業に導入して高校生の英語学習者の自己調整学習と英語力向上に取り組んだ教師の奮闘物語である。私は教師の工夫と生徒たちの努力が意外な効果をもたらしたことに研究上の興味を禁じえない。それは当然の帰結でもあった。まことに貴重な研究成果報告である。

実は指導にあたった私自身，途中でうまくいくのかハラハラした。しかし，最後に得たことは，「勝者はしずかにほほ笑む」であった。

CAN-DOリストとCAN-DO Check Sheetは，本研究の前提として自ら作ってあり，それに慣れていることがうまくいく前提となっていた。それ故に一歩進んで深くつっこんで，指導効果と測定の予想がつき，測定ができた。これを第一の効果としなければならない。

この研究者はベテランである。この2つの道具を利用して，生徒の「自己批判の厳しさ」を発見し，それを打ち破る教師の指導の努力を引き出したのである。一学期には，成績が落ち，生徒の気概が落ちたが，指導者はその理由を理解し，生徒に自信をつけることに成功した。生徒の素直な言葉とそれをかえって利用し，指導力を発揮している。この論文はともすれば，平凡な，または厳しい評価を受けても仕方ないものであった。それを指導力を発揮して，生徒にやる気を奮い起こさせ，能力を持ち上げることに成功した。これこそ教師の教師力である。

私はこの論文から，教師の真剣な努力が生徒の実力を引き出す例を読ませてもらうことができた。気持ちのよい喜びである。

B. 実践部門・報告Ⅲ

和田　稔

ルーブリックを用いて段階を踏んだ、継続的なスピーキング活動の実践

【研究者：鶴田 京子】

本研究の特徴は次の点にある。

①「話すこと」の指導にルーブリックを活用したこと

②指導は「帯活動」を使い，段階的に行ったこと

③「話すこと」の指導を「発表」から「やり取り」へと

発展させたこと
④指導の成果を多角的に検証したこと
一般的には，ルーブリックは教師が指導の結果を確かめるために生徒を対象に行うものだが，この研究では，ルーブリックを生徒と共有することで指導に活用している。この視点は極めて興味深いもので，貴重な情報を提供している。

いわゆる「帯活動」は、最近、注目を集めているが，「帯活動」を行ううえで大切なことは，何故「帯活動」を行うのか，何をどのような順序で行うか，つまり，シラバスに関わる課題である。加えて，「帯活動」の短い学習時間で学習の質を確保できるか，という重い課題がある。本研究では，「発表」についても，「やり取り」についてもシラバスが作成されている（表3及び表4の「指導手順」を参照）。この点に関しては、「指導手順」の学習内容を考えると「帯活動」（10分）で効果的な指導が可能かどうか，について分析・検討が行われ，その結果の報告があるとさらに良かったのではないだろうか。

「話すこと」の指導を「発表」から「やり取り」へと段階的に移行してゆくという発想は興味深い。本来，「発表」と「やり取り」とは異なる言語活動であるが，話題を同一にするとか，質疑応答を工夫するなどして，移行を円滑にしようとする発想は参考になるだろう。

指導の成果の分析・検討が多角的である点，学習者の多様性を十分に意識している点，研究者の配慮の細やかさが見て取れる。

いずれにしても，高等学校での英語指導が直面する課題に挑戦した実践であり，参考になることが多い研究である。

B. 実践部門・報告Ⅳ

池田　央

高校生を対象にした「書くこと」に焦点をあてた協働学習が、英文エッセイにおけるCEFRレベル別語彙使用に与える効果

【研究者：三好 徹明】

この研究目的は表題のとおりだが，より詳しく述べると学生同士のコミュニケーション授業（参加者120名、高校1年生）で，夏休みの課題として英文のエッセイを書かせ，それを他の5名の生徒に読ませてコメントを貰う。その結果をもとに秋から冬にかけて，そのエッセイの書き直しをしてもらう。そのとき使用語彙がどの程度増えているかを見る。すると学期末成績の上・中・下のグループいずれにおいても使用語彙数が増加しており協働学習の効果が見えているようだ。使用した統計分析も慎重で，優れた研究論文となっていて研究者の努力が表れている論文である。

ただ，そのためには背景資料となる生徒の協働学習そのものとCEFRの資料があることが必要なわけで，それが貴重な材料となっている。最近日本ではAL（Active Learning）の存在が話題になっているが，現実にはまだあまり多くはなく，その中でもこの協働学習という形は貴重なものだと思われる。次の実施と研究が望まれているであろう。期待される研究の一つである。

C. 調査部門・報告Ⅰ

吉田 研作

LFCにおけるcore項目の学習：日本人英語学習者に関する音響分析

【研究者：北川　彩】

本稿は，いわゆるLingua Franca English のもっとも大きなテーマである，Lingua Franca Englishのcore となる発音について音響音声学的に検証を加えたものである。日本人にとって難しいといわれる発音のうち，母語の影響によるものを除いた発音を対象とし，特に初級学習者，中級者学習者の発音を中心に分析している。また，出てきた結果についてどのような発音指導を行うと良いかという点について言及している。その中で示唆的なのは，難しい発音を集中的に指導することの是非について，難しい発音は結果として習得に時間がかかるので，時間をかけて継続的に指導することの大切さを提案している。

Lingua Franca Englishのcore となる発音については様々な意見が今までにもあったが，本研究の結果から言えることは，Longが提唱するinteraction 仮説に見られるように，実際のコミ

ュニケーションの過程で，相手と「意味の交渉」をしていく中で，互いにとってもっとも理解可能な発音や文法，また単語が適切インプットになっていく，ということだろう。まずはコミュニケーションを通して相手に理解可能な言語形態を調整し，その上で，どうしてもコミュニケーションの障害となる発音等の訓練をするのが最も有効な方法だといえるだろう。つまり，最初に発音などの形式的正確さを身につけてからコミュニケーション活動に入るのではなく，正確さはコミュニケーションの結果，あるいはその過程で身につくものだということを示唆しているように思われるのである。

C. 調査部門・報告Ⅱ

小池 生夫

発話の諸側面に対する意識の質問紙尺度の開発と妥当性の検証：発話生成モデルの観点から

【研究者：鈴木 駿吾】

本研究は日本人大学生の英語能力の深さに関して，彼らが発話する時の意識に加えて，意識下の構造を包括的に把握し，真の英語力の構造を捉えようとした重層的な研究である。しかも，それを被験者の一人一人に，質問紙調査を行い，記述された文の中から様々な分析を行うことを目指したところに特色がある。ここで質問内容を開発することによって，その妥当性がどのくらいあるかを検証するために精密な理論を作ったところに意義がある。その理論は，第2言語習得研究の発達を踏まえたものである。

本研究者は，この第2言語習得理論の発達がもたらした総合的な発話分析を利用したもので，その発達の歴史を簡潔に述べており，それを読むだけでSLA研究の流れがわかる。英語で質問に応じる106名の被験者（学生）を一人づつ，相当の時間をかけて質問に答えてもらう中で，発話の概念化，言語形成，調音の順序で発話生成する等のSLA理論を使用した。それを実際に被験者の発話を分析することで，一般的な一次要因と個人の特殊要因の差を明らかにして，発話の特徴が個人に応じて異なることを解明した。その一般性と個人性の特徴を分析するためには，一人一人の実験の特徴を分析しなければならない。このような分析は今後，個々の例を積み上げていかなければならない。

本論文は欧米人にも通じるものであるらしい，日本人の特徴などが，この種の分析を行うことによって明らかになるべきものである。このような分析が，今後の最先端の分析方法として英語教育研究に利用される日が近づきつつあるものと言えよう。極めて優れた示唆に富む論文である。

● **第30回「英検」研究助成は下記の日程で行われました。**

項　目	期　間
募集期間	2017年2月～4月7日
選　考	4月15日～6月17日
助成金贈呈式	7月22日
研究期間	7月～2018年5月上旬
報告書提出	2018年5月7日

第30回 研究助成

A 研究部門 ● 報告Ⅰ ● 英語能力テストに関する研究

英文読解におけるモニタリング能力の測定

― 一貫性と結束性の比較を通して ―

研究者：茨城県／筑波大学大学院　在籍　小木曽 智子

《研究助言者：吉田 研作》

概要　本研究では，テキストの言語的なつながりである「結束性」と，テキストの内容的なつながりである「一貫性」という2つのテキスト特性に焦点を当てた。2つの調査により，英検の過去問長文テキストの特性を明らかにし，そうした特性を学習者がどの程度モニターしているかを検証した。

調査1では，英検1級から3級までの長文読解問題で使用されるテキストを対象に，テキストの特性（e.g.，結束性）やテキストの読みやすさが受験級によってどのように異なっているかを分析した。結果より，級が上位になるほど，結束性が低くなっていることが示された。調査2では，学習者がテキスト読解時にテキストの特性（i.e.，結束性，一貫性）に気づいているかどうかを，結束性・一貫性の高い／低いテキストを読んだ際の，テキスト特性判断課題の評定値から調査した。結果より，読み手は，結束性はモニターしていないものの，一貫性に関してはモニターしていることが示された。また，一貫性のモニターに関しては，読み手の熟達度によって，読解時の処理が異なることが思考発話の質的分析から示唆された。

1 はじめに

英文を読む際には，テキストで明示的に提示される情報を理解するだけでなく，テキスト内の情報を関連づけることで，一貫した（coherent）心内モデルを構築することが重要である。一貫した文章の表象を構築することは，読解におけるゴールともされており（e.g., Kintsch, 1988），豊かな文章理解のためには不可欠であるとされる。

テキストの一貫した理解を達成するためには，文章中の個々の要素（e.g.，単語，句，文）の結びつき（i.e.，結束性）を確立し，文章全体の心的表象における整合性（i.e.，一貫性）をとる必要がある（川﨑, 2014）。文章読解におけるテキストの結束性と一貫性については，様々な定義が存在しているが，結束性はテキスト自体の言語的特徴を，一貫性は文章の意味的な心的表象の特徴を指すと考えられてきた（e.g., McNamara, Graesser, McCarthy, & Cai, 2014）。こうした文章の特性に着目した読解研究は主に母語話者を対象に行われて来たが（e.g., McNamara et al., 2014），第二言語として英語を学ぶ学習者を対象とした研究は限られて来た。

本研究では，テキストの言語的側面である結束性と，内容的な側面である一貫性に着目し，読み手が読解時にどれだけテキストの結束性と一貫性に注意を払っているのか，またそのモニタリング力が読み手の熟達度とどのように関係しているかを，日本人英語学習者を対象に検証する。

2 先行研究

2.1 読解とは

テキストの一貫した理解のためには，文章によって提示された情報を自分の既存の知識などと結び付け，心的表象（mental representation）を形成することが必要である（e.g., Kintsch, 1994; van Dijk & Kintsch, 1983）。心的表象とは読み手の心内に保存されるテキスト内容のことを指す。読解により構築される心的表象には，テキストの字義的理解から，読み手の知識と統合された理解レベルまでの3段階：(i) 表層構造（surface structure），(ii) テキストベース（textbase），(iii) 状況モデル（situation model）があり，状況モデルが最も深い理解であると言われている。

読解の目的は，豊かな状況モデルを構築することで一貫した理解を達成することである。そのためには，読み手は語や文の意味理解を達成することに加え，推論や連想，理解のモニタリングといった処理（i.e., 上位処理）を行うことが求められる。しかし，読み手が一度にできる処理は限られているため，特に外国語として英語を学習している読み手にとって，読解中に様々な処理を行うことは難しいと言われている。

2.2 英文読解における結束性

先述のとおり，読解において重要な心的表象の構築のためには，テキストで提示される単語や句といった要素の結合や関係性が確立されることが必要である（川崎, 2014）。こうしたテキスト中の要素の結びつきや明示度合いを表す概念として，テキストの「結束性（cohesion）」が挙げられる。テキストの結束性は“the degree to which the concepts, ideas, and relations within a text are explicit”（O'reilly & McNamara, 2007）”と定義され，テキスト内の情報を重複させたり，文中の要素の関係性を明示するような言語的なシグナル（e.g., 接続詞）を挿入したりすることで，文や文のまとまり，段落，チャプター間で操作することができると言われている（e.g., Kintsch, 1998）。

また，テキストの結束性はテキスト難易度の重要な側面の一つであると言われており，読み手の知識や読解スキルによって結束性が与える影響が異なることが先行研究で示されている。また，先行研究ではテキストの結束性が読み手の文章理解や読解処理に影響を与えるということが示されている（Hosoda, 2016; Ogiso, 2018; O'reilly & McNamara, 2007; McNamara & Kintsch, 1996）。

このように読解に大きく影響を与える結束性について，それを客観的に測定する手法がMcNamaraらによって開発された。McNamaraらはCoh-Metrix computational tool（Graesser, McNamara, Louwerse, & Cai, 2004）という言語指標を自動で算出するツールを開発した。Coh-Metrixは品詞や統語分析，潜在意味解析（Latent Semantic Analysis）といった言語学で広く使用されている要素を使用し，言語的な指標を算出することができるツールである（McNamara, Graesser, & Louwerse, 2012）。

2.3 英文読解における一貫性

結束性はテキストの明示的な言語的特性である一方で，「一貫性（coherence）」は読み手によって構築された心的表象をもとに決められると言われている（e.g., Graesser, McNamara, & Louwerse, 2003）。そのため，テキストの言語的な特性である結束性はテキストの意味的な一貫性に寄与するものの，結束性が高くとも必ずしも一貫性が高くなるわけではないと言われている（e.g., Traxler, 2011）。

文中における一貫性は“the way in which text elements, such as sentences or phrases, belong together in terms of their meaning”と定義される（Warren, 2012）。テキストの一貫性については命題分析（e.g., Kintsch & van Dijk, 1978）やテキストで記述される因果関係のネットワークを分析（e.g., Trabasso & van den Broek, 1985）することによってこれまで評価されてきた。また，テキストの読み手がどのように一貫性を捉えるか

については，「テキストが一貫している（make sense）かどうか」や，「テキストがどの程度一貫しているか」を一貫性判断課題により測定したり（Helder, Van Leijenhorst, & van den Broek, 2016；Nahatame, 2017; Todaro, Millis, & Dandotkar, 2010），一貫していないテキストを読んだときに読み手の読解時間や眼球運動が異なるかどうかを検証したりする（e.g., Albrecht & O’Brien 1993; Ushiro et al., 2018, Wassenberg et al., 2015）ことで調査されてきた。

読み手がテキストの一貫性に敏感かどうか（i.e., テキスト中の意味的な矛盾に気づくかどうか）は，読み手が読解中に一貫性の基準（standards of coherence）を運用しているかどうかを示すとも言われている。読み手がきちんと一貫性の基準を運用して読解している場合には，テキストから読み取れる情報が一貫していない場合に，それに気づくと想定される。そして，読み手が一貫性に敏感である場合には，より精緻に，一貫してテキストを理解していると言える。しかし，読み手の一貫性の知覚は，テキストの要因や読み手の個人差である認知処理傾向などによって変化するとも言われており（Todaro et al., 2010），研究間で一致した結果が得られていない。

2.4 理解モニタリングの重要性

読み手のテキスト理解に関連すると考えられるテキスト要因として，ここまでで述べてきた結束性や一貫性が挙げられる。一方，読解の成否を左右する読み手側の要因として，「理解モニタリング」が挙げられる。理解モニタリングとは，“読み手が自身のテキスト理解を評価するプロセス”で，読解における重要なスキルの1つとされている（van der Schoot, Reijntjes, & van Lieshout, 2012）。読解中の理解をモニターできる読み手は，文中の語の意味がわからない時や，背景知識と内容が合致しない時，情報の統合が難しい時（e.g., テキストが矛盾を含んでいる時）にそれを認識し，うまくいかない場合には適切で補完的な読解行動（e.g., 読み戻り，質問，推論生成）を取ると言われている（Language and Reading Research Consortium, & Yeomans-Maldonado, 2017）。

こうした読み手のモニタリングを測る手法として，従来の研究では矛盾パラダイム（inconsistency paradigm）を用い，矛盾を含むテキストにおける自己ペース読みの読解時間の増加や，先述の一貫性判断課題，読解時の情報への読み戻りといった眼球運動の測定が用いられてきた。先行研究では，モニタリング能力の乏しい読み手は，テキストの一貫性をモニターすることが困難で，矛盾が含まれたテキストの矛盾を検知することができないことが示されている（e.g., Albrecht & O’Brien, 1993）。特に日本人英語学習者を対象に，一貫性へのモニタリングを検証した先行研究（Ushiro et al., 2018）では，読み手は常にではないが，一貫性をモニタリングすることができることもあるということが示唆されている。

2.5 本研究の概要と目的

ここまで，読解におけるテキスト要因としての結束性と一貫性についてその定義を確認し，それに関連する理解モニタリングの研究をみてきた。本研究では外国語としての英語学習者である日本人を対象に，テキストの結束性と一貫性に対する読み手のモニタリング能力を比較・検証する。まず調査1では，現行の英検で使用されている長文問題のテキスト特性を確認する。続く調査2では，日本人英語学習者を対象に，学習者が実際に英文読解においてテキスト特性をモニターしているかどうかを検証する。

3 調査1

3.1 目的

調査1では，英検で使用されている各級の長文読解問題のテキスト特性を，結束性，一貫性を中心とする観点から数値化することで，各受験級における英文の実態について明らかにすることを目的とする。具体的には，1級から3級までのテキストを対象とし，異なる難易度の級に

おいてどのようなテキスト特性が見られるかを明らかにする。検証課題(Research Questions: RQs)は次のとおりである:

RQ1-1 Coh-Metrixにより測定される5つのテキスト特性について, 英検1級から3級の長文テキストの現状はどのようになっているか。

RQ1-2 算出されたテキスト特性は, 統計的に有意に各級で異なっているか。

3.2 方法

3.2.1 英検テキストの収集

英検長文読解問題で使用される, 英文の特性を明らかにするために, 英検1級から3級までの長文テキストを収集した。具体的には, 1998年第1回から2016年第2回までの試験で使用された, 物語文・説明文・評論文を可能な限り収集した。なお, ジャンルの違いによる影響を排除するため, 電子メールや掲示といった長文や, 穴埋め式問題として提示されている長文は除外した。分析対象とした各級のテキストの平均語数と読みやすさの指標であるFlesch-Kincaid Grade Levelは, 表1のとおりである。

3.2.2 分析で使用する指標

本調査では, 長文全体の外観をつかむため, テキスト特性に関する指標をMcNamara et al. (2014)より選定した。実際に, 分析に使用した指標は次の5つである:Narrativity, Syntactic Simplicity, Word Concreteness, Reference Cohesion, Deep Cohesion。上記の5つの指標は, 0から100の値をとり, 値が大きいほど読みやすいテキストであることを意味する。

この指標についてさらに具体的に説明する。Narrativityはテキストがどれだけ物語文的であるかを示すもので, 値が大きいほど読み手にとって馴染み深いことを示している。Syntactic Simplicityは, テキストに含まれる文がどれだけ少ない単語・シンプルで読み手にとって馴染み深い文構造を用いているかを示し, 値が大きいほど統語的に複雑でないことを表すものである。Word Concretenessは, テキストがどれほど具象性の高い単語を含んでいるかを示すものである。具象性とは, 読み手にとってその意味内容が明確で, イメージしやすいかどうかを表すものである。このWord Concretenessの値が大きいほど, 読み手にとって具体的な語彙が使われていることが確認できる。Reference Cohesionは, テキストに含まれる概念や出来事のオーバーラップによってテキストがどれだけ明示的かを示す値である。この値が大きいほど, 読み手にとって明示的で結束性が高いテキストであることが確認できる。Deep Cohesionは, テキスト内にどれだけ因果や論理的な関係性が示されているかを示す値である。この値が大きいほど, 読み手は一貫した理解を構築しやすいと言われている。 本研究では上述の5つの指標を用いて, 英検で過去に使用された長文の特徴を捉えることを試みた。

3.3 手順

英検の問題で使用された長文テキストを, ウェブ上のテキスト特性算出ツールであるCoh-metrix3.0 (http://www.cohmetrix.com/) にかけることで, テキスト特性の値を算出した。

■表1: 調査1で分析したテキストの語数と読みやすさ

	1級 (*n* = 160)		準1級 (*n* = 132)		2級 (*n* = 97)		準2級 (*n* = 66)		3級 (*n* = 46)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
語数	547.19	147.93	414.52	73.31	357.73	9.33	291.65	7.89	253.13	5.88
FKGL	12.95	1.90	12.11	9.60	19.96	1.16	18.97	0.99	12.15	1.01

(注)*n* = テキストの数, FKGL = Flesch-Kincaid Grade Level.

3.4 結果と考察

表2に，Coh-Metrixで算出したテキスト特性を級ごとに集計した結果を示す。また，図1に級ごとのテキスト特性の値を可視化したものを示す。記述統計値から，Narrativityについては，下位の級の方になるほど値が大きくなっており，下位の級になるほど読み手にとって馴染み深い長文が使用されている傾向にあることが読み取れる。Syntactic SimplicityとWord Concretenessについても同様の傾向が見られ，下位の級であるほど読み手にとってシンプルでわかりやすい構造で，イメージしやすい具体的な単語が使用されていることが読み取れる。

続けて，本研究の焦点である結束性・一貫性に関連する指標であるReference Cohesion，Deep Cohesionについても見てみる。記述統計値より，上位級であるほど値が小さく，下位級であるほどその値が大きくなっていることが示された。このことから，英検で使用されているテキストは受験級が上位になるほど結束性・一貫性が低くなる傾向にあり，上位級ほど読み手にとって難易度の高い英文になっていることが確認された。

■表2: 実験1における各級のテキスト特性

	1級 (n = 160)		準1級 (n = 132)		2級 (n = 97)		準2級 (n = 66)		3級 (n = 46)	
特性	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
N	21.58	12.13	21.83	9.60	36.86	13.46	49.53	15.55	56.57	18.24
SS	37.06	14.53	38.21	14.16	48.10	14.43	53.89	15.48	70.59	13.12
WC	51.13	20.49	61.45	22.22	60.63	23.78	79.23	16.51	87.30	11.76
RC	16.74	14.26	19.73	16.32	44.95	20.76	56.74	19.98	68.22	18.50
DC	63.75	18.94	64.16	21.60	74.64	19.53	77.97	20.75	66.72	23.58

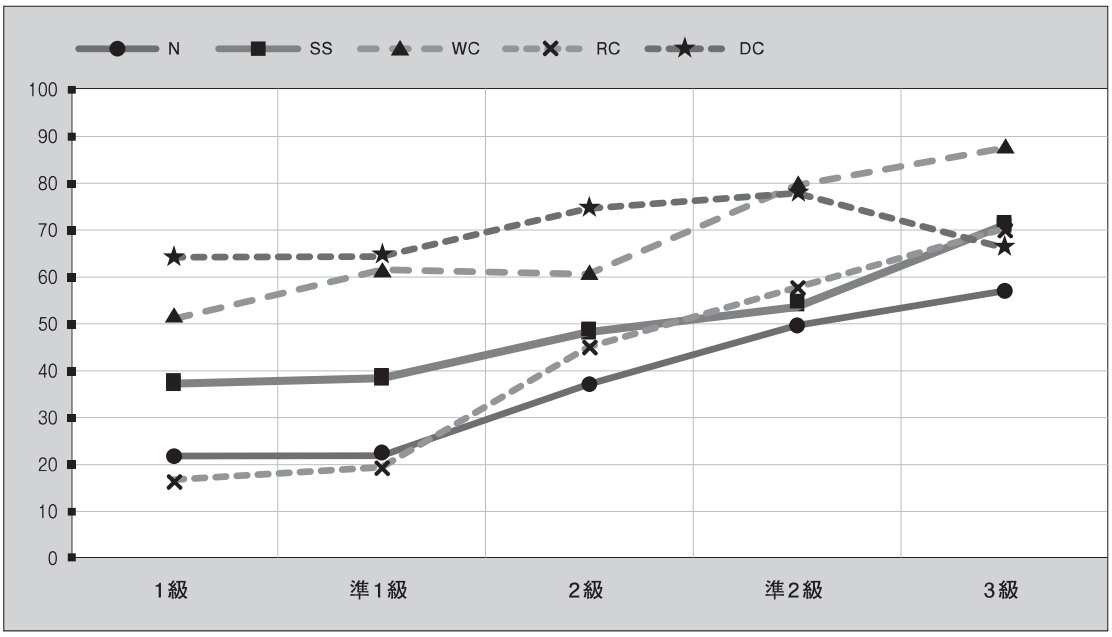
(注)n = テキストの数, N = Narrativity, SS = Syntactic Simplicity, WC = Word Concreteness, RC = Reference Cohesion, DC = Deep Cohesion.

続いて，これらのテキスト特性について，級間で統計的に有意に異なっているかを確認するため，テキスト特性 [Narrativity, Syntactic Simplicity, Word Concreteness, Reference Cohesion, Deep Cohesion] ×英検級 [1, 準1, 2, 準2, 3] の二元配置の分散分析を行なった。テキスト特性を対応あり要因に，英検の級を対応なし要因として分析を行った。分散分析の結果，テキスト特性の主効果，英検級の主効果，テキスト特性×英検級の交互作用の全てにおいて，統計的な有意差が確認された($p < .001$)。

結果より，英検で過去に長文問題として使用されたテキストの特性は級ごとに異なっており，難易度が高くなるほど，テキストの読みにくさが増すこと，すなわち，読み手にとって馴染みのないトピック(i.e., Narrativity)や複雑な文構造(i.e., Syntactic Simplicity)，イメージしにくい抽象的な語(i.e., Word Concreteness)が使用され，テキストの結束性(i.e., Reference Cohesion)や意味的な一貫性(i.e., Deep Cohesion)が低くなることが示された(RQ-1, RQ-2)。

3.5 調査1のまとめ

調査1では，これまでに1級から3級までの英検の筆記試験で使用された長文問題を，Coh-Metrixにより分析し，算出されたテキスト特性を統計的に分析した。結果から，英検の受験級が高くなるほど，英文の容易さを示すテキスト特性(e.g., 結束性や一貫性)の値が低くなり，テキストが読みにくくなっていることが示唆された。続く調査2では，読み手がテキストを読んでいる際に，このようなテキスト特性を実際にモニターしているかを調査するため，テキスト特性判断課題を実施した。



■図1: 級ごとのテキスト特性の推移, 扱ったテキスト特性は以下のとおり: N = Narrativity, SS = Syntactic Simplicity, WC = Word Concreteness, RC = Reference Cohesion, DC = Deep Cohesion.

4 調査2

4.1 目的

調査2では，日本人英語学習者が実際にテキスト特性（結束性，一貫性）をどれだけモニターしているかを調査する。検証課題（RQs）は以下のとおりである：

RQ2-1 異なる英語熟達度の学習者は，テキストの結束性を正しく捉えることができるか。

RQ2-2 異なる英語熟達度の学習者は，テキストの一貫性を正しく捉えることができるか。

RQ2-3 モニタリングの際に学習者はどのような読解処理を行なっているか。

4.2 方法

4.2.1 協力者

調査2に参加した協力者は，国立大学・大学院に通う30名の日本人英語学習者であった。協力者の年齢は19歳から27歳で，男性が21名，女性は9名であった。協力者の専攻は，人文学，教育学，社会学，理工学と多岐に渡った。協力者は全員日本語母語話者で，少なくとも6年以上英語を学んでいた。自己申告の協力者の英語関連保有スコアによると，協力者の英語熟達度レベルは英検4級～準1級，TOEIC L & R 520点から970点程度であり，本研究の協力者はCEFR A1からC1レベルの英語学習者であると推定された。

4.2.2 マテリアル

(1) テキスト

説明文読解に関する先行研究より，科学的な現象を説明する短いパッセージを選定した。結束性に関するパッセージとして，日本人英語学習者を対象とした説明文読解研究（Hosoda, 2016; Ogiso, 2018）より，2つの説明文（Heat Changes Matter [以降Heat], Space Travel and the Human Body [Space]）が選定された。テキ

ストの結束性の操作はOzuru, Dempsey, and McNamara(2009)のテキスト操作方法に基づき,主に以下の4つの観点から行われた:(1)曖昧な代名詞を名詞に置き換える,(2)馴染みのない概念と馴染みのある概念をつなぐ詳細な記述の追加,(3)文同士やアイデア間の関係性を明確にするための接続詞の挿入,(4)隣接する文間のオーバーラップを増やすための語の置き換えや挿入。結束性に関するテキスト例は表3に示すとおりである。

一貫性に関するパッセージには,一貫性の矛盾のモニタリングを検証した先行研究(Steffens, 2016)より,2つのパッセージ(How Bees Make Honey [以降Bee], How Hiccups Occur [Hiccup])が選定された。パッセージは以下の文章構造であった:はじめにトピックに関する導入がIntroductionでなされ,後にEvent sentenceでターゲットとなる因果的な出来事についての記述がなされ,Target sentenceで因果関係の一致・不一致に関係する一文が提示され,End of chainで文章が締めくくられた。一貫性の高い・低い条件を設けるにあたり,Target sentenceで説明されていた出来事と概念的に一致するEvent sentences中の一語をそのまま(一貫性高)あるいは反意語(一貫性低)にすることで操作した。表4に,パッセージBeeのテキストを例として示す。なお,一貫性に関しては,Steffens(2016)において,一貫性の高・低バージョンで,意味的な差異があることは確認されていた。

■表3: 結束性を操作したマテリアルHeat

Adding heat or taking away heat can change matter. Matter can change from one state to another state, or from one form to another form. Three states of matter are solid, liquid and gas. **For example,** an ice cube is solid water. Heat can melt an ice cube, causing the ice cube to change into liquid water. When heat is taken away, the liquid water can change back into solid water. Heat can make liquids boil and change into a gas state. **For example,** water boils when it is heated. As the water boils, it turns into a gas state that is called water vapor. Heat from the sun causes liquid water to turn into water vapor. Water vapor **then** mixes with the air in a process called evaporation. **However,** sometimes heat causes irreversible changes. **As one example,** bread can change into toast when you heat the bread. **However,** you cannot untoast a piece of toast by taking away heat. **As another example,** eggs change when you cook them in a pan. **However, of course** you cannot uncook an egg by taking away the heat.

(注)太字,下線,二重下線の修正は高結束性条件で提示された:太字 = 接続詞等の言語シグナル,
下線 = 代名詞化や省略が可能な指示・照応表現他,二重下線 = 付加情報.

■表4: 一貫性を操作したマテリアルBee

構造	構造
Introduction/ start of chain	Bees are like humans in that they make their own food, honey. The process of honey production starts by worker bees seeking out flowers. The bees find flowers and begin to drink their liquid nectar. The nectar is stored in a special stomach, separate from where food is stored, called the honey stomach. Enzymes in this special stomach break down the nectar into sugars and water. This broken down nectar and water are regurgitated into the honey cell of the hive. The bees begin to move their muscles.
Event sentences	The movement of the wing muscles **warms**/**cools** the surrounding air and the honey cell.
Target sentence	The **heated** water inside of the cell evaporates.
End of chain	The sugars inside the cell thicken, forming honey.

(注) Event sentenceにおける太字の単語warmsは一貫性高条件, coolsは一貫性低条件で提示され, それぞれTarget sentenceの太字heatedと一貫する/しないよう操作された.

■表5: 実験パッセージの特徴

テキスト特性	一貫性		一貫性	
テキスト名	Bee	Hiccup	Heat	Space
語数	118	130	157	128
FKGL	6.3	7.8	3.7	5.3
FRE	70.8	59.4	81.0	77.3

(注) FKGL = Flesch-Kincaid Grade Level, FRE = Flesch Reading Ease, なお上記の数値は一貫性高条件, 低結束性条件のパッセージの値である.

なお, 上記のテキストを選定する際には, 英語教育を専攻する大学院生2名で, テキストのトピックや使用語彙について協議を行った。また, 本調査の対象者が外国語として英語を学んでいる読み手であることを考慮して, JACET8000 (JACET, 2003)の5以上の単語は, 易しい単語やフレーズで言い換えるか, 注釈によって日本語でその意味を付与した。各テキストの語数や読みやすさについては, 表5に示すとおりである。

(2) 思考発話タスク

協力者が読解中にテキストの一貫性の矛盾や, テキストの結びつき(結束性)の弱さをモニターし, それに対して読みを変化させているかといった処理を調べるために,「読解中に頭に浮かんだことを発話する」という思考発話タスクを行った。協力者は本番のテキストを読む前に, 思考発話の説明を受け, 実際に思考発話をしているモデル音声を聞くことでタスクに関して理解を深めた。また, 短い練習用テキストを用いて, 実際に

考えていることを発話する練習を事前に行った。疑問がある場合には，本番前に実験者に聞く機会が設けられ，タスクに関する問題が解消された。本番の協力者の思考発話はICレコーダーにて録音され，音声は書き起こしたデータとして使用された。

(3) テキスト特性判断課題

協力者がテキスト特性を正しく捉えているかどうかを確かめるために，テキスト特性についての判断課題を行なった。テキスト内容について「筋が通っていたか(make sense)」を問うために，先行研究で用いられていた一貫性判断課題(consistency judgement: Helder, Van Leijenhorst, & van den Broek, 2016; Nahatame, 2017)を参考に，結束性・一貫性に関する判断課題を作成した。テキストの言語的なつながりである結束性に関しては，「先ほど読んだ英文は，内容がはっきりと，わかりやすく関連づけられて記述されていた」かどうか，テキストの内容的なつながりである一貫性に関しては，「先ほど読んだ英文は，内容にまとまりがあり，筋が通って，一貫していた」かどうかを，協力者にたずねた。協力者は，「1：全く思わない」から「6：とてもそう思う」までの6段階の回答尺度の中から，最も当てはまると思うものを一つ選ぶように指示された。この判断課題は，各テキストを読み終わるごとに協力者に評定してもらった。

(4) 読解熟達度テスト

協力者の読解熟達度を測るため，過去に使用された英検の準1級，2級テストから，5つの長文大問を選定し，読解熟達度テストを作成した。テストには，多肢選択式問題が24項目含まれており，一項目1点，合計24点満点のテストとして使用した。実施時には，30分の制限時間を設けた。

4.2.3 手順

調査は個別に実施され，所要時間は約80分であった。実験の流れについては，図2に示すとおりである。はじめに，協力者は30分間で読解熟達度テストに取り組んだ。5分ほどの休憩の後，読解中に考えていることを話すという思考発話タスクの説明とモデルCDを聞き，実際に練習用パッセージに対して思考発話を行うことで，協力者は活動の練習を行った。思考発話に関して疑問点があった場合は，ここで解決された。

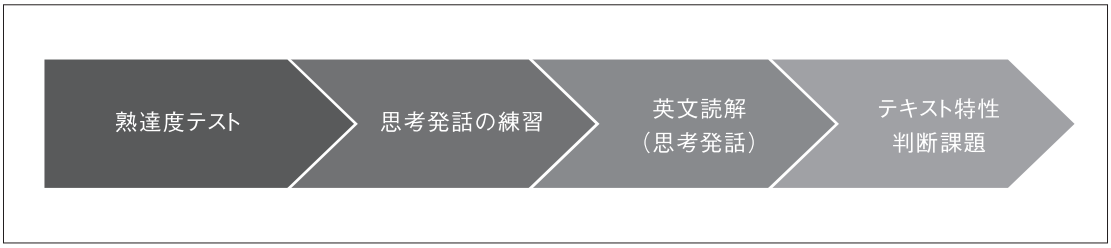
その後，協力者は，「内容を理解するため」に英文を読むように伝えられ，実験テキストを読解し，思考発話タスクを行った。この時，一貫性に関するテキストは一方が一貫性の低いバージョン，他方は一貫性の高いバージョンで提示された。また，結束性に関するテキストについても同様に，一方が低結束性バージョンで，他方は高結束性バージョンで提示された。テキストの提示順は，協力者ごとにランダマイズされた。

そして，協力者は，テキストを読み終えるごとに一貫性・結束性判断課題に回答した。なお，テキスト読解後に判断課題に回答することは，協力者に事前に伝えられていた。この読解活動は平均して30分程度であった。

4.2.4 採点・分析

(1) 思考発話タスク

協力者のモニタリングや読解中の処理を見るため，協力者の発話を書き起こした。なお，今回はテキスト特性判断課題が主な分析であるため，読解中の読み手の処理内容を見るための質的データとして考察に使用した。



■図2: 実験2の調査手順

(2) テキスト特性判断課題

評定値(1-6)を集計し，二元配置の分散分析により，一貫性や結束性の条件間で，読み手の評定値に違いがあるかどうかを調査した。

(3) 読解熟達度テスト

多肢選択式の一項目を1点とし，全24項目24点満点としてテストを採点した。

4.3 結果と考察

(1) 読解熟達度テスト

熟達度テストの記述統計を表6に示す。読解熟達度テストの信頼性はCronbach's α = .74であった（注．24項目で算出した信頼性が.71と低かったため，他の項目に比べて熟達度下位者が不自然に正答できていた1項目をテストから削除した)。得点の中央値に基づき，協力者(N = 30)を熟達度上位群(n = 19)と下位群(n = 11)の2群に分割した。熟達度上位群と下位群の間の熟達度テストの点数に，統計的に有意な違いがあることを確かめるため，t検定を用いて分析した。結果より，上位群と下位群の読解熟達度テストの得点には，統計的に有意な差があることが確認された，$t(28) = 6.72$, p = .017, r = .79。

■表6: 読解熟達度テストの記述統計

熟達度	*n*	*M*	95%CI	*SD*	*Min*	*Max*
上位群	19	15.68	[14.74,16.36]	1.79	14.00	20.00
下位群	11	10.09	[8.41,11.77]	2.84	5.00	13.00
合計	30	13.63	[12.38,14.88]	3.49	5.00	20.00

(2) テキスト特性判断課題

■結束性判断課題

結束性判断課題の記述統計を表7に示す。記述統計量より，概して，協力者は結束性の高いテキストにおいて，結束性の低いテキストよりも高い評定値をつけていたことがわかった。

■表7: 結束性判断課題の記述統計

		結束性高			結束性低		
熟達度	*n*	*M*	95%CI	*SD*	*M*	95%CI	*SD*
上位群	19	4.84	[4.44, 5.25]	0.90	4.68	[4.26, 5.11]	0.95
下位群	11	4.64	[3.79, 5.48]	1.43	4.36	[3.76, 4.97]	1.03
合計	30	4.77	[4.37, 5.16]	1.10	4.57	[4.22, 4.91]	0.97

熟達度上位群と下位群において，結束性の異なるテキストでの評定に差があるかどうかを検証するため，テキストの結束性×協力者の熟達度の二元配置分散分析を行なった。テキストの結束性を協力者内要因に，協力者の熟達度を協力者間要因とした。分散分析の結果，結束性の主効果，$F(1, 28) = 0.71$, p = .405, 熟達度の主効果，$F(1, 28) = 0.74$, p = .400, 結束性×熟達度の交互作用，$F(1, 28) = 0.05$, p = .823, について，統計的に有意な効果は確認できなかった。

このことから，結束性判断課題の結果によると，読み手は，テキストの結束性には敏感でない(i.e., 結束性のモニタリング能力がない)可能性が示された(RQ2-1)。これは，読み手はテキストの字義的な明示性やテキスト自体の構造については，読解中に注意を払っておらず，テキストの意味的な理解に注力していたためであると考えられる。また，今回用いた結束性判断課題は，先行研究(e.g., Helder et al., 2016)に基づいて実験者がオリジナルで作成したものであったため，課題自体の妥当性が低かった可能性も考えられる。

■一貫性判断課題

続いて，一貫性判断課題の記述統計を表8に示す。記述統計より，協力者は一貫性の高いテキストにおいて，一貫性の低いテキストよりも高い評定値をつけていたことが示された。

■表8: 一貫性判断課題の記述統計

		一貫性高			一貫性低		
熟達度	*n*	*M*	95%CI	*SD*	*M*	95%CI	*SD*
上位群	19	4.74	[4.25, 5.23]	1.10	4.47	[3.94, 5.00]	1.17
下位群	11	4.73	[4.20, 5.26]	0.91	3.73	[2.81, 4.65]	1.56
合計	30	4.74	[4.37, 5.09]	1.02	4.30	[3.72, 4.68]	1.35

熟達度上位群と下位群において，一貫性の異なるテキストでの評定に差があるかどうかを検証するため，テキストの一貫性×協力者の熟達度の二元配置分散分析を行なった。テキストの一貫性（高/低）を協力者内に，協力者の熟達度（上位群/下位群）を協力者間要因とした。分散分析の結果，一貫性の主効果が有意であった，$F(1, 28) = 5.05$, $p = .033$。しかし，熟達度の主効果と一貫性×熟達度の交互作用については統計的に有意な差は見られなかった。それぞれの統計値は以下のとおりであった，$F(1, 28) = 1.16$, $p = .290$, $F(1, 28) = 1.72$, $p = .201$。

このことから，読み手はテキストの一貫性に敏感であり，実際にモニタリングしていることが示唆された（RQ2-2）。読み手の熟達度による差異は確認されなかったが，今回対象とした協力者は国立大学に通う大学生・大学院生であったため，ある程度の英語熟達度を有していたと考えられる。そのため，ある程度の英語熟達度を越える学習者であれば，テキストの内容的なつながりである一貫性に注意を払って読解し，テキストの一貫性を左右する矛盾を検知することができると考えられる。読み手が一貫性をモニターすることができるという結果は，日本人英語学習者を対象とした先行研究（e.g., Ushiro et al., 2018）とも一致するものであった。また，本研究において読み手がモニタリングに成功した理由として，テキスト自体の分量が少なく，一貫性の操作としてテキストに挿入された矛盾は，因果関係が破綻する，共存し得ない明確な矛盾であったためであると考察できる。

■テキスト特性判断課題の比較

続いて，読み手の一貫性・結束性のモニタリング力に違いがあるかどうかを検討するため，それぞれのテキストにおける判断課題の評定値（1-6）が「[一貫性/結束性] 高条件 > [一貫性/結束性] 低条件」となった協力者を“モニタリングができた人”，条件間で評定値が同じあるいは「[一貫性/結束性] 高条件 < [一貫性/結束性] 低条件」となった協力者を“モニタリングできなかった人”とラベル付けした。各ラベルにおける人数を算出し，以下のクロス集計表（表9）にまとめた。

カイ二乗検定の結果，$\chi^2 = 0.60$, $p = .438$ で一貫性と結束性の間のモニタリングの成否に関して，統計的な有意差は確認できなかった。この結果に関しては，前述の結束性に関してモニタリ

■表9: 各テキストをモニタリングできた人数に関するクロス集計表

テキスト特性	モニタリングできた人数	モニタリングできなかった人数	合計
結束性	13	17	30
一貫性	16	14	30
合計	29	31	60

ングが見られなかったこと(RQ2-2)も影響している可能性がある。本研究では,テキストの言語的な特性である結束性と,テキストの意味的な特性である一貫性を比較することを試みたが,両者のモニタリングに関しては,テキスト特性間で有意な差が見られないことが示された。

(3)思考発話タスク

続いて,熟達度の異なる読み手が実際にどのような処理を行っていたかについて,質的に分析するために,これまでの分析結果より,読み手がモニタリングすることができていた「一貫性」に焦点を当てて,読み手の発話を記述した。矛盾のあった一貫性低テキストBeeにおける,協力者の発話例を表10に示す。なお,読み手の熟達度を考慮するため,読解熟達度テストにおいて上位だった協力者と,下位だった協力者の発話を記述した。

■表10: 一貫性操作テキストにおける協力者の発話例

テキスト原文	The bees begin to move their muscles. The movement of the wing muscles **cools** the surrounding air and the honey cell. The **heated** water inside of the cell evaporates.
熟達度上位者	で蜂はその筋肉を動かし始めて…その羽の奥の筋肉の動き…というか羽の動きで周囲の筋肉を**冷やし**…そして同時にhoney cellも**冷やしている**…おっと coolしていたのにheated water inside of the cell…あれそういうことかどういうことだ…外の空気は**冷やされる**んだけど体内の水は**体温が上がる**ってこと…すごいな
熟達度下位者	これらの力を動かして…動き…羽の動き…冷たく囲む…空気じゃない…空気と蜂蜜の部屋を…保つってことか…蒸気になる…**あっためられた水**は蒸気になる

(注)太字 = 矛盾 (i.e., coolとheated) に関連する箇所,下線部 = 協力者が一貫性をモニタリングして矛盾に気づいた箇所.

Beeテキストでは,働き蜂がどのように蜂蜜を作るかが記述されている。矛盾テキストでは,『蜂が羽を動かして周囲の空気を「温め」ることで,「熱された」水が蒸発する』という箇所について,『空気を「冷やす」』という矛盾が含まれている。熟達度上位者は,「羽の動きで周囲の筋肉を冷やし…そして同時にhoney cellも冷やしている」とまず文の意味を確認した後に,「おっと coolしていたのにheated water inside of the cell…あれそういうことかどういうことだ」と,テキストに含まれている一貫性の矛盾に気づいている。その上で,「外の空気は冷やされるんだけど体内の水は体温が上がるってこと…すごいな」と,テキスト情報の矛盾を,統合することで解決し(Hakala & O'Brien, 1995),テキスト情報を理解していることがわかる。

一方,熟達度下位者は,表10に示されているように,常に文章の字義的な処理に従事しており,矛盾箇所には気づいていないことがわかる。今回のテキスト特性判断課題の結果では,一貫性のモニタリングについて,協力者の熟達度に関連したモニタリング能力の差異は確認できなかったが,熟達度の違いによって読解中の処理の仕方や,矛盾に気づいた際の読解方略に違いがあることが示された(RQ2-3)。

4.4 調査2のまとめ

調査2では,日本人大学生・大学院生30名を対象に,テキスト特性判断課題を行うことで,読み手がテキストの結束性と一貫性をモニターしているかどうかを測定した。純粋にテキスト特性に応じて評定できていた協力者は,結束性・一貫性ともに全体の半数程度であった。

各テキスト特性について統計的に分析をしたところ,結束性については,読み手は統計的に有意に,一貫性の高低に応じて評定値を変化させることができていること(i.e., 一貫性のモニタリング)が確認された。一方で,結束性に関しては,読み手はモニタリングできていないことが示唆さ

れた。これは，読み手，とりわけ外国語として英語を学習している読み手は，テキストの意味処理を中心に処理を行っており，テキストの言語的特徴である結束性には敏感でないためであると考察される。

また，テキスト特性判断課題では協力者の熟達度の上位者および下位者間にモニタリング能力の差異を確認することはできなかったが，読解中の発話を質的に見たところ，読解中の処理には違いがある可能性も示唆された。具体的には，矛盾をモニタリングした上で，情報を一貫して理解しようとする方略的な読みが見られた。こうしたことから，読み手のモニタリング能力には，熟達度だけでなく，読み手の他の認知能力や，読解時の方略も関係している可能性が示唆された。

5 結論と今後の課題

本研究では，テキストの言語的なつながりである結束性と，内容的なつながりである一貫性に焦点を当てた調査を2つ行った。調査1では，英検1級から3級の過去問題の長文テキストの特性をCoh-Metrixによって算出した。結果から，級が難しくなるほどテキストが読みにくい，難易度の高いものになっていることが示唆された。また，読解に影響を与えると先行研究で示唆されているテキストの結束性や一貫性に関しても，難易度の高い級において低い値であることが確認された。また，調査2では，調査1でも扱ったテキストの結束性と一貫性に焦点を当て，操作した結束性・一貫性テキストを読んだ際に，それらのテキスト特性の高低を正しく認知し，モニタリングできているかどうかを，テキスト特性判断課題によって測定した。日本人英語学習者を対象に調査をしたところ，読み手はテキストの一貫性にはモニターしているものの，テキストの結束性に関してはモニターしていないことが示唆された。また，読解時の読み手の処理を思考発話法により見たところ，矛盾に気づいていた熟達度の高い読み手は，矛盾していても一貫してテキストを理解しようと，情報を統合することで読解していることが明らかになった。

本研究の結果より得られた教育的示唆としては，以下の点が挙げられる。調査1の結果より，難しいテキストにおいてはその結束性や一貫性の値が低くなることが示された。このことから，教師やテスト・教材作成者は，単なるテキストの古典的な読みやすさ指標（e.g., Flesch-Kincaid Grade Level）だけでなく，文章の結束性や一貫性に注意する必要があることが示唆された。Coh-Metrixのようなウェブツールの存在により，現在は容易にテキスト特性を算出することが可能である。本研究より，読み手に影響を与えうるテキスト特性について，今後より注意を払う必要があるということが示唆された。

さらに，調査2の結果から，読み手は読解において，テキストの意味的な一貫性には注意を払っている一方で，テキストの言語的な構造にはそれほど意識的に注意を向けていないことが明らかとなった。本調査では，読み手はテキストの一貫性に注意を払って読解している，「良い読み手」であることが示された。しかし，熟達度が十分でない場合や，馴染みのない文章である場合には，テキストの構造や接続詞が読解をサポートすると言われている（e.g., McNamara & Kintsch, 1996）。ゆえに，テキスト読解時には，テキストの内容と合わせて，そこで使用されている言語やテキストの構造に着目し，読解時には方略的に読んでいくことを明示的に指導することが，読解に苦手意識を持つ読み手のサポートになると考えられる。

本研究では，読解に影響を与えられると考えられてきたテキストの字義的な特徴である結束性と，内容的な一貫性を操作することで，読み手のモニタリング能力を測定することを試みた。外国語読解研究において，両者のモニタリングを直接的に比較検証したものは非常に限定的であるため，本研究は外国語読解研究に新たな示唆を与えることを試みた，萌芽的な研究であると考えられる。しかし，本研究には留意すべき点や限界点も存在している。調査2で用いたテキスト特性判断課題は，本来一貫性に関する判断課題（e.g., Helder, et al. 2016）であったものを，実験者が結束性に関しても援用したものであるため，その方法論的妥当性は十分でなかった可能性が考えられる。さらに，調査2の協力者は国立大学生・大学院生

30名であり，実施した読解熟達度テストにおける協力者の得点のばらつきもあまり大きくなかった。そのため，協力者の熟達度も考慮した分析が十分にできなかったという限界点がある。さらに，本研究の主眼ではなかったため，読み手の読解中の処理方略や日常の読解活動で用いているストラテジーを調査した結果を考慮して分析することができなかった。

今後の研究では，上述の限界点を考慮し，より幅広い学習者を対象に調査を行うことで，読解中の処理（e.g.，思考発話のプロトコル）や，日常のストラテジー使用も要因に含めた分析・考察をすることが望まれる。さらに，テキストの結束性と一貫性のモニタリングに関する，より良い測定手法を開発することで，読み手のモニタリング能力と読解力の関係性について示唆を与えることが望まれる。

謝辞

本研究を実施・発表する貴重な機会を与えて下さいました，公益財団法人 日本英語検定協会の皆さま，ならびに選考委員の先生方に心より御礼申し上げます。特に，助言者である吉田研作先生には，有益なご助言・ご指導をいただきましたこと，深く感謝申し上げます。さらに，筑波大学大学院の卯城祐司先生には，本研究の立案から実施，報告書執筆に至るまで，親身なご指導をいただきました。心より感謝申し上げます。また，細田雅也さん，森好紳さんをはじめとする，研究室の先輩・同輩・後輩の皆さまにも多くのご助言と励ましをいただきましたこと，深く御礼申し上げます。最後に，本調査に協力して下さった皆さまに厚く御礼申し上げます。

参考文献（＊は引用文献）

＊ Albrecht, J. E., & O' Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1061-1070.

＊ Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking Reading Comprehension,* 82-98.

＊ Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36,* 193-202.

＊ Hakala, C. M., & O'Brien, E. J. (1995). Strategies for resolving coherence breaks in reading. Discourse Process, 20, 167-185.

＊ Helder, A., Van Leijenhorst, L., & van den Broek, P. (2016). Coherence monitoring by good and poor comprehenders in elementary school: Comparing offline and online measures. *Learning and Individual Differences, 48,* 17-23.

＊ Hosoda, M. (2016). The interplay of text cohesion and L2 reading proficiency in different levels of text comprehension among EFL readers. *ARELE: Annual Review of English Language Education in Japan, 27,* 201-216.

＊ JACET. (2003). *JACET list of 8000 basic words.* Tokyo, Japan: Author.

＊ 川崎恵里子（編）. (2014).『文章理解の認知心理学ーことば・からだ・悩』. 東京：誠信書房.

＊ Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163-182.

＊ Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49,* 294-303.

＊ Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

＊ Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

＊ Language and Reading Research Consortium, & Yeomans-Maldonado, G. (2017). Development of comprehension monitoring in beginner readers. *Reading and Writing, 30,* 2039-2067.

＊ McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

＊ McNamara, D.S., Graesser, A., & Louwerse, M. (2012). Ch.6 Sources of text difficulty: Across genres and grades. In Sabatini, J., Albro, E, & O' Reilly, T. (2012) *Measuring up: Advances in how we assess reading ability* (pp. 89-116). Lanham, MD: Rowman 6 Little field Education.

＊ McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22,* 247-287.

＊ Nahatame, S. (2017). Standards of coherence in second language reading: Sentence connectivity and reading proficiency. *Reading in a Foreign Language, 29,* 86-112.

＊ Ogiso, T. (2018). Effects of text cohesion on inference generation during EFL reading: Evidence from think-aloud protocols. *ARELE: Annual Review of English Language Education in Japan, 29,* 161-176.

＊ O'reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes, 43,* 121-152.

＊ Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19,* 228-242.

＊ Steffens, B. (2016). *Detecting coherence breaks while reading scientific explanations* (Doctoral dissertation). Northern Illinois University, United States of America.

＊ Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24,* 612-630.

＊ Todaro, S., Millis, K., & Dandotkar, S. (2010). The impact of semantic and causal relatedness and reading skill on standards of coherence. *Discourse Processes, 47,* 421-446.

＊ Traxler, M. J. (2011). *Introduction to psycholinguistics: Understanding language science.* Hoboken, NJ: John Wiley & Blackwell.

＊ Ushiro, Y., Hosoda, M., Nahatame, S., Mori, Y., Suzuki, K., Tada, G., Ogiso, T., Kamimura, K., Sasaki, Y., & Mandokoro, R. (2018). Understanding Protagonist, Causal, and Intentional Links During EFL Narrative Reading. *ARELE: Annual Review of English Language Education in Japan, 29,* 81-96.

＊ van der Schoot, M., Reijntjes, A., & van Lieshout, E. C. (2012). How do children deal with inconsistencies in text? An eye fixation and self-paced reading study in good and poor reading comprehenders. *Reading and Writing, 25,* 1665-1690.

＊ van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York, NY: Academic Press.

＊ Warren, P. (2012). *Introducing psycholinguistics.* New York, NY: Cambridge University Press.

＊ Wassenburg, S. I., Bos, L. T., de Koning, B. B., & van der Schoot, M. (2015). Effects of an inconsistency-detection training aimed at improving comprehension monitoring in primary school children. *Discourse Processes, 52,* 463-488.

第30回 研究助成
A 研究部門 ● 報告Ⅱ ● 英語能力テストに関する研究

多肢選択の解答パターンに基づく未知語推測プロセス診断：形態素・文脈情報の活用から

研究者：茨城県／筑波大学大学院　在籍・日本学術振興会特別研究員　神村 幸蔵
《研究助言者：村木 英治》

概要

本研究では，英語学習者の未知語推測中に利用される情報に焦点を当て，解答パターンから学習者の従事する未知語推測プロセスを判断することができる多肢選択式テストの作成について示唆を得ることを目的に調査を行った。具体的には，選択肢の内容に目標語中の形態素，特に接頭辞の意味が反映されているもの（形態素情報）と選択肢の内容が問題文の文脈に合致するもの（文脈情報），これらの情報を掛け合わせたもの，そしてこれらの情報がまったく利用できない選択肢を持つ多肢選択式問題を作成した。学習者が目標語を推測した後に選んだ選択肢によって学習者が未知語の意味推測に抱える困難を解明することを目的とした。先行研究の知見を踏まえ，手がかりの使用に影響を与える要因として学習者の語彙サイズと英文読解熟達度を考慮した。結果から，形態素・文脈情報の利用可能性，多肢選択式未知語推測テストの選択肢の選ばれる傾向および学習者の語彙サイズの間に関係性が明らかとなった。また，本研究の結果から，多肢選択式未知語推測テスト作成と教育現場に対する示唆が得られた。

1 はじめに

語彙知識は読解をはじめ第二言語（second language; L2）の運用において重要な要素の1つである（Schmitt, 2008）。そのため，学習者は大量の語彙を覚える必要がある。語彙知識の広さ（学習者が知っている英単語の数；語彙サイズともいう）と英文読解の能力の関係はこれまでの研究によって検証されてきた。Nation（2006）は，英語学習者が教師や辞書の助けを借りて英文を読むことができるようになるためには，約4,000ワードファミリー（Grabe [2009] の換算によると約10,000見出し語）が必要であると述べている。さらに，英語学習者がサポートなく英文を流暢に読むためには，テキストの98%以上の語を知っている必要があり，このテキストに占める既知語の割合を達成するためには9,000ワードファミリーが必要である（Nation, 2006）。このことから，流暢な英文読解を達成するために，学習者は相当量の英単語を学ばなければならないということがわかる。

しかしながら，学習者がそのような大量の語彙をすべて単語帳による暗記や教師の指導による意図的な学習（意図的語彙学習）によって習得することは，現実的ではない（Grabe, 2009）。そのため，学習者はしばしば，読解中に未習語または既知語でも意味を知らない単語（未知語）と出会うことがある。その際，学習者は出会った未知語の意味を推測すること（未知語推測）によって流暢にテキストを読み進めたり，付随的に語彙を学んだりすること（付随的語彙学習）が求められる。しかし，未知語推測は失敗に終わることが多いストラテジーでもある。学習者の未知語推測を検証したNassaji（2003）では，意味を正しく推測できたのは25.6%で，部分的に成功したのも

18.6%であったことが報告されている。したがって，学習者が流暢な読解や付随的語彙学習を行うために，未知語の意味の推測方法を明示的に指導し，その能力を伸ばす必要があると考えられる。実際，日本の英語教育の文脈においても，現行の高等学校学習指導要領では，コミュニケーション英語Ⅱの授業において「未知の語の意味を推測したり背景となる知識を活用したりしながら聞いたり読んだりすること」(文部科学省，2009，p. 111)を指導することが定められている。

以上のことから，学習者の読解中の未知語推測のプロセスの解明や，その能力の向上に関してさらなる検証が求められることがわかる。学習者の未知語推測能力を向上させるためには，教師の指導だけでなく学習者の未知語推測能力を測定するためのテストも必要となる。本研究では，未知語推測中に利用される情報に焦点を当て，解答パターンから学習者の従事する未知語推測プロセスを判断することができる多肢選択式テストの作成について示唆を得ることを目的に調査を行った。

2 先行研究

2.1 読解における未知語推測の重要性とそのメカニズム

学習者が読解中に未知語と出会った際に採りうるストラテジーには，未知語を無視する，辞書を引いて意味を調べる，未知語の意味を推測するということがあげられる(Ender, 2014)。このうち，未知語を無視して読み進めるストラテジーは，形容詞や副詞といった品詞の単語に対して効果的であるとされる。これは，それらの品詞の単語を飛ばして読むことが文の内容の理解の妨げにならないためである(Aebersold & Field, 1997)。また，辞書を引くことは，語彙学習に効果的であることが検証されている(Cho & Krashen, 1994)。しかし，単語を無視し読み飛ばすストラテジーは，名詞や動詞といった文理解において比較的重要な品詞の語に対しては必ずしも有用であるとは限らない。また，辞書を引くことは，目標語(未知語)の見出しを探すのに時間がかかるため(Knight, 1994)，学習者の流暢な読みを阻害する恐れがある。これらのストラテジーのうち，未知語推測は，最も頻繁に行われ，かつ望ましいストラテジーの1つである(Schmitt, 2010)。実際に，Paribakht and Wesche(1999)によって，学習者は読解中に出会った未知語の70%以上の単語に対して意味の推測を行うことが報告されている。また，未知語推測は，付随的語彙学習のみならず(Nation, 2013)，テキスト理解においても重要である(Zhang & Koda, 2012)ことが検証されている。このことから，読解中の未知語推測は，学習者の語彙学習や読解において重要な役割を果たしているといえる。

未知語推測(lexical inferencing)とは，先行研究(Haastrup, 1991)において次のように定義されている：

> The process of lexical inferencing involves making informed guesses as to the meaning of a word in light of all available linguistic cues in combination with the learner's general knowledge of the world and her awareness of the co-text and her relevant linguistic knowledge (p. 40).

すなわち，学習者は，一般的な知識とともに未知語が含まれる文脈，そして関連する言語の知識といった利用可能な手がかりを組み合わせて，読解中に出会った未知語の意味を推測すると考えられる。

Haastrup(1991)を含め，先行研究において，L2の未知語推測における学習者の手がかりの利用と推測の処理プロセスに関するモデルが提案されている。たとえば，Huckin and Bloch(1993)は学習者が利用可能な手がかりを用い

て自らの推測の仮説を検証するモデルを提案した。このモデルでは，読解中に学習者のワーキングメモリに入ってきた未知語の手がかりの情報によって仮説生成の試行が始まる。その後，自らの持っている知識（e.g.，形態素，テキスト内容，一般知識）をもとに仮説を評価する。また，de Bot, Paribakht, and Wesche（1997）は，LevelIt（1993）の単語処理のモデルに基づいて未知語推測の処理プロセスを説明している。LevelIt（1993）の単語処理のモデルでは，心的辞書がすべての言語処理の中央に位置しており，語彙知識の貯蔵庫のような役割を果たしているとされる。心的辞書は見出し語と語彙素からなる。見出し語は語の意味と統語情報を含んでおり，語彙素には形態的・音韻的情報が含まれている。de Bot et al.は，L2の未知語推測とは，上述の心的辞書の空の見出し語に本来あるべき情報を埋めようとすることであると説明している。具体的には，学習者が未知語に遭遇したとき，はじめに未知語の語形と語彙素に保存されている情報を照応させようとする。その際，意味を持つ形態素情報や文の情報，一般知識が用いられ，学習者は自らの既知の概念と結びつけようとする（de Bot et al., 1997）。これらの研究から，学習者は未知語推測を行う際，(a)学習者が利用可能な手がかりを組み合わせ，(b)その手がかりを用いて未知語の意味に関する仮説を検証し，(c)未知語を学習者の既知の概念と結び付けるというプロセスに従事すると考えられる。本研究では，これらのプロセスのうち，初期段階にあたる利用可能な手がかりを組み合わせるプロセスに焦点を当て調査を行う。

2.2 未知語推測に使用される手がかり

上述のとおり，これまでの研究によって提案されてきたモデルでは，未知語推測の処理プロセスの1つに学習者が利用可能な手がかりを用いる段階があると考えられている。英語学習者の未知語推測における手がかりの利用に関しては，すでにいくつかの研究において実際に検証されている。たとえば，英語学習者の大学生を対象に付随的語彙習得の調査をしたParibakht and Wesche（1999）では，学習者が未知語の意味を推測する際に用いた情報源について検証した。結果として，文レベルの文法的知識（e.g.，語順による語の品詞特定）が35%，語の形態素（e.g.，語幹や*-tion*, *-ly*といった接辞）が15%，固有名詞の語頭の大文字や複数の同列のものを列挙する際に使用されるコンマなどの句読法（punctuation）の情報が11%の割合で使用されていたことが明らかとなった（表1）。

■表1: 未知語推測の際に学習者の用いた知識源の割合 (%)(Paribakht & Wesche, 1999)

文レベルでの文法知識	形態素に関する知識	句読法	世界知識	同音異義語	談話レベルの知識	連想	同語源	不明
35	15	11	9	4	4	3	2	18

また，Nassaji（2003）では，第二言語としての英語を学ぶ学習者を対象に未知語推測中に用いられる知識源の違いを検証した（表2）。その結果，推測中に使用される知識源のうち，世界知識（world knowledge; 一般常識的な知識）が最もよく用いられ（46, 2%），ついで形態素の知識が多く用いられていた（26.9%）。その他にも，文法の知識（11.5%）や談話レベルの知識（8.7%）といった，文や文章の手がかりも用いられていたことが明らかとなった。

■表2: 未知語推測の際に学習者の用いた知識源の割合 (%)(Nassaji, 2003)

世界知識	形態素知識	文法知識	談話レベルの知識	L1の知識
46.2	26.9	11.5	8.7	6.7

これらの研究から，英語学習者の未知語推測において，世界知識や形態素の知識，文や談話レベルの知識がよく用いられることが示されている。本研究では，これらの手がかりのうち，テスト問題や選択肢によって比較的情報が操作しやすい形態素の情報と文脈の情報に着目して調査を行った。

2.2.1 未知語推測における形態素情報の活用

大部分の英単語は複数の意味をもつ最小の構成要素（i.e.，形態素）から成り立っている。たとえば，*blackboard* や *unhappy* はそれぞれ，*black*（黒い）と *board*（板），*un-*（否定）と *happy*（幸せな）とに分けられる。前者の *black*，*board* や後者の *happy* のように，ある語を構成している形態素それ自体が語として数えられるものを自由形態素（独立形態素）といい，*unhappy* の *un-* のようにそれ自体では語として存在することができないものを拘束形態素という。これらの形態素の意味を分析し組み合わせることによって語全体の意味（e.g.，「黒板」，「不幸せな」）を明らかにすることができる。このように語を分析し個々の構成要素の意味を結びつけることでその単語の全体の意味を知る能力が未知語推測において重要な役割を果たすということが，英語を母語（first language; L1）とする話者と（e.g.，McCutchen & Logan, 2011） L2として英語を学ぶ学習者を対象に行った研究（Zhang & Koda, 2012）それぞれにおいて示されている。このことから，未知語中の形態素の情報を意味の推測に役立てることは、L1の話者のみが行う難易度の高いストラテジーではなく，L2の学習者にも利用可能なものであると考えられる。

2.2.2 未知語推測における文脈情報の活用

形態素情報と同様に，語を含む文脈の情報も未知語の意味推測においては重要であり，学習者の文脈情報の利用についてもさまざまな研究によって検証されている。その1つにWebb（2008）があげられる。Webbは，未知語の意味推測と付随的語彙学習における文脈の情報の効果について調べることを目的として日本人EFL学習者を対象に調査を実施した。具体的には、目標語の意味を推測できる情報の量が異なる文脈を提示し，その文脈に含まれる疑似語について意味の再生と再認のテスト（多肢選択式）を行った。この研究では，目標語の意味を推測するための情報量について英語母語話者に予備調査を行い，文脈が推測するための情報をどれくらい含んでいるのかを評価させた。実験結果より，未知語の意味推測に役立つ情報を含んでいる文脈を学習者が読んだ場合、情報を含んでいない文脈に比べ意味の再生において有意に高かったということが明らかとなった。このことから，英語学習者の未知語の意味推測お

■表3: Hamada (2014) の未知語推測タスクの例

形態素の条件	問題と選択肢の例
MR条件	No one had an umbrella. We stood under the rainfime. a. covered tent b. rain shelter c. candle light d. ling grass e. I don't know
MU条件	The student is starting college this semester. She is buying a rainfime. a. new laptop b. rain shower c. long river d. cold mountain e. I don't know

（注）目標語には下線が引かれた。

よびその保持に対し，文脈の情報の効果があることが示された。

2.2.3 未知語推測における形態素・文脈情報の活用

英語学習者の未知語推測における形態素情報と文脈情報の利用について調査した研究にHamada(2014)がある。この研究では，未知語の意味推測をする際に目標語中の形態素情報(自由形態素)に頼ることのできる条件(morphology reliable条件; MR条件)と形態素情報に頼ることのできない条件（morphology unreliable条件; MU条件)を設定し，学習者の英語の熟達度(ACT Compassの結果により分けられた)によってMR条件とMU条件のときに学習者が利用する2種類の情報（i.e., 形態素情報、文脈情報)の違いを検証した。検証は、表3のような多肢選択式による未知語推測タスクによって行われた。結果として，熟達度に関わらずMR条件においては形態素の情報を利用して未知語推測に役立てられたこと，そして，MU条件において熟達度の高い学習者は文脈の情報を利用し未知語の意味推測に成功したのに対し，低から中レベルの熟達度の学習者は形態素の情報に頼ってしまい，誤った意味を推測してしまうことが示された。この研究では、実在語と疑似語を組み合わせた疑似的な複合語を目標語として使用しており，目標語中の実在語の部分を自由形態素として扱っていたが，構造的・書記法的に複雑な派生接辞(e.g., *pre-*)について扱う必要性について言及している。また，限界点の1つとして，形態素の情報を統制したにも関わらず文脈の情報については統制することができなかった点をあげている。

また，同様に，英語学習者の未知語推測に関して目標語の語形(i.e., 形態素)の情報と文脈の手がかりの両方を取り扱った研究としてNakagawa(2006)がある。この実験では，2つのグループに分けた協力者へ2つの未知語推測テストを課した。まず，未知語推測テストⅠとして，あるグループには語形から目標語の意味を推測するテスト(Word-Based Test: WBT)，もう一方のグループには文脈の情報から目標語の意味を推測するテスト(Context-Based Test: CBT)を実施した(表4)。そして，両方のグループに対して未知語テストⅡとして両方の情報が与えられたテストを行い，未知語の品詞と意味の特定に形態素情報と文脈の手がかりのどちらが利用可能であるかについて協力者の熟達度別に検証した。

■表4: Nakagawa (2006) で使用された未知語推測テストの例

テストの種類		テスト形式の例
Lexical Inference Test I	WBT	Recourse
	CBT	The police officer tried to get the demonstrators to stop blocking traffic, but they refused. In the end, the only recourse was to arrest them.
Lexical Inference Test II		The police officer tried to get the demonstrators to stop blocking traffic, but they refused. In the end, the only recourse was to arrest them.

(注)目標語には下線が引かれた。WBT = Word-Based Test; CBT = Context-Based Test。

結果として，未知語の意味を特定するうえで文脈の手がかりのほうが語形の情報に比べて有効であることが明らかとなった。ただし，この研究では，目的語となる語の形態素を学習者が利用可能か，つまり知っていたかどうかが考慮されていなかった。たとえば，この実験で扱われた目標語に*recourse*という語がある。今回の実験においてこの単語は「頼りとする人」という意味で用いられている。しかし，この語を形態素である*re-*(再び，後ろに)と*course*(進路，方向)の意味を組み合わせても「頼りとする人」という意味にはならない。Nakagawaの実験結果は，この

ような形態素の情報に頼ることができない語を目標語に含んでおり，語形の情報のみを与えられた学習者は形態素の情報を利用することができなかったため、文脈の情報が利用可能であった条件に比べ意味の推測ができなかったと考えられる。また，学習者が推測した目標語の意味のうち、3語分について載せているが，*repast* や *unwind*、*prevaricated* についてそれぞれ「もう一度」，「無風の」，「前に壁を作る」というように各語の接頭辞の意味を推測した意味に反映していることがうかがえる。このことから，英語学習者は、目標語の中の形態素情報，とりわけ接頭辞が意味の推測をする際に頼ることのできる場合，語形から未知語の推測をすることが可能であることが考えられる。

形態素情報を用いた未知語推測と読解力の関係を検証した研究にZhang and Koda(2012)がある。この研究では，構造方程式モデリングを用いて，単語を形態素へ分割する能力が未知語推測の能力を介して，(a)英語学習者の語彙知識に寄与すること，さらに(b)語彙知識を介して読解力へ貢献することが示された。

2.3 本研究の枠組み

先行研究より，未知語推測には(a)学習者が利用可能な手がかりを組み合わせ，(b)その手がかりを用いて未知語の意味に関する仮説を検証し，(c)未知語を学習者の既知の概念を結び付けるという処理プロセスが関わると仮定される(de Bot et al., 1997; Huckin & Bloch, 1993)。その中で，学習者の使用する手がかりには，一般的な知識のほかに，未知語の形態素や未知語の周囲の文や談話といった文脈の情報が用いられていることが明らかとなっている(Nassaji, 2003; Paribakht & Wesche, 1999)。また，その手がかりの利用は学習者の語彙サイズや読解力といった英語熟達度と関係していることも示されている(Hamada, 2014; Zhang & Koda, 2012)。

以上のことを踏まえ，本研究では，形態素・文脈情報の活用という未知語推測の処理プロセスに着目し，多肢選択式の未知語推測テスト作成に有益な示唆を得ることを目的に調査を行った。具体的には，選択肢の内容に目標語中の形態素，特に接頭辞の意味が反映されているもの(形態素情報)と選択肢の内容が問題文の文脈に合致するもの(文脈情報)，これらの情報を掛け合わせたもの，そしてこれらの情報がまったく利用できない選択肢を持つ多肢選択式問題を作成した。作成された未知語推測テストの選択肢のうち，学習者が目標語を推測した後にどの選択肢を選んだかによって学習者が未知語の意味推測に抱える困難を解明することを目的とした。 先行研究の知見を踏まえ，今回は手がかりの使用に影響を与える要因として学習者の語彙サイズと英文読解熟達度を考慮した。

本研究で取り組んだ研究課題(research questions; RQs)は以下の通りである。

RQ1 英語学習者は，形態素情報と文脈情報を掛け合わせた多肢選択式未知語推測テストに解答する際，目標語の形態素情報や目標文の文脈情報の手がかりを用いて意味を推測するか？

RQ2 英語学習者の語彙サイズテストと英文読解熟達度テストの成績は，形態素情報と文脈情報を掛け合わせた多肢選択式未知語推測テストの選択肢の選択率を予測するか？

3 方法

3.1 協力者

本調査に参加した協力者は茨城県の私立大学に通う大学生113名(実験実施時の年齢: 18-23歳; 平均)であった。協力者は，ネパール出身の1名を除き，日本語母語話者であり，日本国内の公教育において最低でも6年間の英語教育を受けていた。協力者の専攻は，看護学，英語学，心理福祉学であった。本調査の目的および内容を説明した後，協力者から同意を得た。しかし，本調査で実施する複数のテストのうち，1つでも受験しなかった協力者のデータや各テストのうち外れ値と判断したデータを除外したため，最終的に108名のデー

タが分析対象となった。

協力者のおおよその英語熟達度を把握するために，協力者が有している大規模英語テストのスコアや級を紙面にて尋ねた。その結果，74名が英検の取得級（協力者の取得級：4〜2級，平均=3級〜準2級，4級=6名，3級=19名，準2級=32名，2級=17名）を報告し，1名がTOEIC Listening & Readingのスコア（735）を報告し，37名はどの資格も報告しなかった。なお，複数のスコア・級を報告した者はいなかった。報告されたスコア・級を「外国語の学習，教授，評価のためのヨーロッパ共通参照枠（Common European Framework of Reference for Languages: Learning, teaching, assessment; CEFR）」と照らし合わせた結果，本調査の協力者は初級-中級（A1:「基礎段階の言語使用者」- B1:「自立した言語使用者」）に相当する英語熟達度を有することが分かった。

3.2 実験材料

3.2.1 語彙サイズテスト

協力者の語彙サイズを測定するために，筆記版の望月語彙サイズテスト（相澤・望月，2010; 以下語彙サイズテスト）を用いた（資料1）。これは，望月（1998）によって開発されたテストであり，日本人のテスト受験者の推定される語彙サイズを7,000語まで測定することが可能である。このテストは，1,000語レベルから7,000語レベルまでを1,000語ごとに7つのレベルに区切られている。各レベルで日本語の対訳となる26語の英単語を選択していく形式のテストとなっている。今回は第2版の語彙サイズテストのうち1,000語レベルから7,000語レベルの7つのレベルの問題を用い，協力者の語彙サイズを7,000語レベルまでの範囲で測定した。

3.2.2 英文読解熟達度テスト

協力者を対象に英語で書かれた長文の読解力を測定するため，英文読解熟達度テスト（読解熟達度テスト）を作成し，使用した。このテストの作成には，公益財団法人日本英語検定協会が実施する標準化テストである実用英語技能検定（英検）の過去問題（旺文社，2015ab, 2017ab）を採用した。これは次の2つの理由による。第一に，英検が日本国内において広く実施されている英語熟達度測定テストであり，その出題形式が多くの日本人にとってなじみがあると考えられるためである。第二に，英検は級ごとに難易度が異なるため，協力者のレベルに応じた難易度の問題が設定しやすいためである。

本調査では，読解熟達度測定に英検2級と準1級から3つずつ，合計6つの多肢選択式の長文読解問題を採用した（合計27問；2級：15問，準1級：12問）。これは，英検2級が高等学校卒業レベルであることと（日本英語検定協会，2018），本調査の協力者が大学生であることを考慮し，2級と準1級の問題を用いることが英文読解熟達度の弁別に妥当であると考えたためである。

3.2.3 未知語推測テスト

協力者の未知語推測の処理過程を確かめるために，先行研究（Kamimura, 2017）で用いられた未知語推測タスクをもとに紙面の未知語推測テストを作成し，用いた。目標語として，Kamimura（2017）にて用いられた英単語22語を採用した（資料2）。各目標語は，接頭辞＋自由形態素（単語）から成り，推測に有益な情報がある / ない文脈上において提示された（文脈の有益性）。文脈の有益性については，Kamimura（2017）の予備調査において確認されている。

本研究の関心は，未知語推測の処理プロセスを多肢選択の解答パターンから見ることができるかであるため，先行研究（Hamada, 2014）を参考に，上述の目標語・文脈を含む多肢選択式問題を作成した（資料3）。各多肢選択式の問には，下線が引かれた目標語を含む文と4つの英語で書かれた選択肢，そして，目標語が協力者にとって既知語であるかどうかを把握するために，「この単語を知っていましたか？」という問を設けた。選択肢は，目標語中の形態素の情報と文脈の情報の有無を掛け合わせて作成した。資料3の問（目標語は*biplane*［複葉機］）を例に説明すると，目標語に含まれる接頭辞の*bi-*の持つ意味（i.e.,「2つ」,「複数」）を反映させた選択肢（i.e., 4番の選択肢；Mor選択肢とよぶ），文脈には合致する（目標語と代替して不自然でない）が目標語の意味ではない選択肢（i.e., 1番の選択肢；Con選択肢とよぶ），形態素の意味を含み文脈にも合致する選択肢（2番の選択肢i.e., MorCon選択肢とよぶ），そして

目標語の形態素情報とも文脈情報とも合致しない選択肢（3番の選択肢；Non選択肢とよぶ）の4種類の選択肢を作成した。すべての問いにおいて，MorCon選択肢が正答であった。文脈の有益性の条件間のカウンターバランスを取るために，セットAとBの2種類の未知語推測テストを作成した。したがって，22の目標語にそれぞれ2つずつ，合計44文を問題文として使用した。

3.2.4 接頭辞テスト

協力者に本研究の目標語が含む接頭辞に関する知識があるかを確かめるために，筆記の接頭辞テストを実施した。本研究では，Kamimura（2017）によって作成された接頭辞テストを採用した。L2学習者の接頭辞知識を測定している先行研究（Mochizuki & Aizawa, 2000）では，再認課題によって接頭辞の知識を確かめている。しかし，形態素情報から未知語の意味を推測するためには，語形を見て意味を思い出す再生的な接頭辞の知識が必要であると考えられる。そのため，接頭辞に関する再生知識を測定するテストとして用いられたKamimuraの接頭辞テストを本研究において使用した。

今回使用した接頭辞テストは，上述の未知語推測テストの目標語に合わせ，22の問から成る。各問には目標接頭辞と疑似語幹から成る3つの疑似語があり，接頭辞部分には下線が引かれた（資料4）。

3.3 手順

本調査の協力者は，英語の授業中，または個別にテストに取り組んだ。本調査では，協力者に対し，4種類のテストが課された。調査は2週間にまたがって実施された。1週目には，語彙サイズテストと読解熟達度テストが実施され，残り2つのテストは2週目に行われた。調査はすべて，調査者である筆者の指示に沿って行われた。

はじめに，第1週目の最初に語彙サイズテストの冊子が協力者に配布された。この冊子には，1,000語レベルから7,000語レベルまでの7つのセクションがあり，協力者はそれぞれのセクションを3分で解答するように指示された。

つぎに，読解熟達度テストの問題冊子と解答用紙が協力者に配布された。協力者は，英検2級と準1級の長文読解問題6題を解くために30分与えられ，時間配分に気を付けながら可能な限りたくさんの問題に解答するよう指示された。また，今回調査に用いたテキストが協力者にとって既読でないことを確かめるために，協力者は，問題文となっているテキストを今までに読んだことがあるかについて「はい」または「いいえ」の欄に○を付けて回答するよう指示が与えられた。

読解熟達度テストに取り組んだ1週間後に，協力者は，未知語推測テストに取り組んだ。協力者には未知語推測テストの問題冊子セットAまたはBのいずれかがランダムに配布された。協力者は，次の手順に沿って問題を解いた：はじめに，問題文中の下線の引かれている単語（目標語）の意味を推測した。つぎに，目標語が協力者にとって既知でなかったことを把握するために，協力者は，下線が引かれている単語をすでに知っているかどうかを判断し、「この単語を知っていましたか？」という欄の「はい」または「いいえ」のいずれかのボックスにチェックを入れて回答した。最後に，推測した意味に合う選択肢の数字を○で囲った。目標語が既知であるかを尋ねる問に関して，解答開始前に協力者に対し，「知っている単語」とは、語の一部分だけ知っているのではなく、あくまで「語全体の形を知っている，または見たことがある単語」を指すということが伝えられた。これは，協力者が目標語の接頭辞または語幹をすでに知っていることのみをもって「知っている単語」と判断することで，実際には目標語そのものは未知語であるにもかかわらず既知語として分析から除外してしまうことを防ぐためであった。また，先行研究（Hamada, 2014）を参考に，目標語の意味の推測が難しくても可能な限り意味を推測してすべての問いに解答するよう協力者に指示を与えた。時間制限は設けなかった。

最後に，接頭辞テストが協力者に配布された。協力者は，知っている範囲で各問にある3つの疑似語の下線が引かれた接頭辞の意味を解答欄に日本語で記入するよう指示された。同時に，どうしても接頭辞の意味が分からなければ、意味の欄は未記入のままにするよう伝えられた。これは，このテストの目的が，協力者の有している接頭辞の意味の知識を把握することであり，協力者がその場で意味を推測し解答してしまうことを防ぐ

ためであった。これに加え，協力者は，心当たりのある答えが複数ある場合は、解答を3つまで記入するように指示が与えられた。これは，目標となっている接頭辞の中には多義のものが含まれており（e.g., *ex-* には「外に」のほかに，「元の」という意味もある），解答の個数を1つだけに絞ると協力者の持つ知識のうち一部しか見ることができない恐れがあったためである。

3.4 採点と分析

語彙サイズテストの採点は，解答を項目ごとに0（不正解）と1（正解）の2値に分けることにより行われた。これらの2値データを相澤・望月（2010）の示す次の計算式に代入することでテストを受けた協力者の推定語彙サイズを求めた：推定語彙サイズ = 正答数÷26×1000。また，平均＋3標準偏差の範囲よりも小さい値を算出した協力者1名分のデータがあったが，分析を安定させるためにそのデータを分析から除外した。信頼性係数は十分な値であった（Cronbach's α = .94）。

読解熟達度テストの採点は，語彙サイズテストと同様に，0（不正解）と1（正解）の2値に分けることにより行われた。すべての項目を入れた読解熟達度テストの信頼性が低かったため，負の相関を示した項目6つを削除したが，依然として十分な値の信頼性は得られなかった（Cronbach's α = .60）。

未知語推測テストに関して，協力者1人当たりの各選択肢（i.e., MorCon/Mor/Con/Non 選択肢）の選択率を算出した。また，それぞれの選択肢について，接頭辞利用可能／不可能条件（+接頭辞/-接頭辞）×文脈の有益性高／低い（+文脈/-文脈）条件ごとの選択率も同様に算出した。選択率は，当該の選択肢が選ばれた数を協力者1人当たりが解答した項目の数（i.e., 22問）で割ることによって算出した。読解熟達度テストと同様に，このテストの信頼性も十分な値とはいえなかった（Cronbach's α = .36）。

接頭辞テストは，語彙サイズテストと同様に，解答を項目ごとに0（不正解）と1（正解）の2値に分けることにより行われた。3.3節にて述べたとおり，目標となっている接頭辞の中には多義のものが含まれているが，今回は目標語中で用いられているときの接頭辞の意味が書き出されている場合にのみ正解と評価した。例えば，*ex-* には「外に」や「元の」という意味があるが，本調査の目標語である *excommunication*（追放）においては「外に」の意味で用いられているため，「外に」に準ずる解答のみを正解と評価した。正答であった接頭辞を接頭辞利用可能条件の接頭辞，不正解であったものを接頭辞利用不可能条件とした。

RQsに答えるために，未知語推測テストの各選択肢の選択率を従属変数，語彙サイズテストと読解熟達度テストの成績を独立変数とした強制投入法による重回帰分析を行った。今回は，どの選択肢についても1つの独立変数についてのみ有意な結果が得られたため，より予測率の高いモデルを得るために，それぞれの選択肢を従属変数とした強制投入法による回帰分析を行った。

4 結果と考察

4.1 協力者の語彙サイズおよび英文読解力熟達度

語彙サイズテストと読解熟達度テストの結果をまとめたものが表5である。読解熟達度テストに関して，21点満点（信頼性の低い6項目削除済み）で平均点が約5点であり，床面効果がみられた。すなわち，今回の英文読解熟達度テストは，今回の調査の協力者にとって難易度が高かったことが示された。このことは信頼性の低さ（Cronbach's α = .60）からうかがうことができる。

■表5：語彙サイズテストと読解熟達度テストの記述統計

テストの種類	M	95%CI	SD	Min	Max
語彙サイズテスト	3,428.06	[3,261.81 ,3,594.32]	871.57	1,038	5,308
読解熟達度テスト	4.57	[4.10 ,5.05]	2.49	0	13

（注）読解熟達度テストは21点満点であった

4.2 形態素情報と文脈情報の利用可能性と未知語推測テストの選択肢の選択率の関係

RQ1に答えるため，接頭辞利用可能性と文脈の有益性を掛け合わせた各条件における未知語推測テストの選択肢の各種類の選択率をまとめた（表6）。表6から，接頭辞の利用可能性と文脈の有益性を掛け合わせたどの条件においてもMorCon選択肢（形態素の意味を含み文脈にも合致する選択肢）がもっとも多く選ばれていることが明らかとなった。また，Non選択肢（目標語の形態素情報とも文脈情報とも合致しない選択肢）がどの条件においてももっとも選択率が低かったことが示された。これは，学習者が未知語推測を行う際に，利用可能な手がかりを組み合わせるプロセスがあることを示す先行研究（de Bot et al., 1997; Huckin & Bloch, 1993）を支持するものである。本調査の場合，協力者にとって形態素情報（接頭辞情報）と文脈情報が利用可能な手がかりであった。したがって，協力者は，目標語（目標文）の語形や文脈の手がかりを組み合わせるプロセスを経て，意味を推測しようとしていたことが示唆される（RQ1）。

■表6: 未知語推測テストの選択肢ごとの選択率

選択肢の種類	＋接頭辞		-接頭辞	
	＋文脈	-文脈	＋文脈	-文脈
MorCon	.41	.34	.42	.43
Mor	.13	.21	.16	.16
Con	.24	.17	.25	.26
Non	.08	.11	.16	.15

（注）＋接頭辞/-接頭辞：接頭辞利用可能/不可能条件；＋文脈/-文脈：文脈の有益性高/低い条件

4.3 協力者の英語熟達度と未知語推測テストの選択肢の選択率の関係

RQ2に答えるために，協力者の語彙サイズテストおよび読解熟達度テストの成績と未知語推測タスクの各選択肢の選択率に焦点を当てて分析を行った。表7は語彙サイズテスト，読解熟達度テストおよび未知語推測テストの各選択肢の選択率の相関行列をまとめたものである。

相関分析の結果，語彙サイズテストの成績と読解熟達度テストが相関関係にあることがわかった（r = .242, p < .05）。語彙知識の広さが読解力と関係していることは多くの先行研究によって明らかにされてきている（e.g., Zhang & Koda, 2012）。しかし，相関が弱い点は先行研究（e.g., Grabe, 2009）の指摘とは異なる。これは，今回の読解熟達度テストにおいて床面効果が出たためであると考えられる。また，全体的な傾向として，

■表7: 語彙サイズテスト，読解熟達度テストの成績および未知語推測タスクの各選択肢の選択率の相関関係 (N = 108)

選択肢の種類	*M*	*SD*	1	2	3	4	5	6
1. 語彙サイズテスト	3,428.06	871.57	—					
2. 読解熟達度テスト	4.57	2.49	.241*	—				
3. MorCon	0.43	0.13	.476**	.082	—			
4. Mor	0.18	0.10	-.447**	-.18	-.587**	—		
5. Con	0.24	0.09	.112	.243*	-.255**	-.348**	—	
6. Non	0.15	0.09	-.301**	-0.15	-.521**	.075	-.248**	—

（注）*p < .05, **p < .01.

語彙サイズテストの成績と未知語推測テストのConを除くすべての選択肢の選択率との間に相関がみられた(*p*s < .01)。また，読解熟達度テストの成績はCon選択肢と相関関係にあることがわかった(*r* = .243, *p* < .05)。

以上の相関分析によって示された語彙サイズテスト，読解熟達度テストの成績および未知語推測タスクの各選択肢の選択率の相関関係をより詳細に解明するために回帰分析を行った。具体的には，未知語推測テストの各選択肢の選択率を従属変数，語彙サイズテストと読解熟達度テストの成績を独立変数とした強制投入法による重回帰分析を行った。重回帰分析を行う前に，分析の前提となる多重共線性の確認を行ったところ，どの変数間にも *r* が .80を上回る相関関係はなかったため，多重共線性が無かったとみなした。

重回帰分析の結果，投入された2要因での回帰モデルがすべての選択肢について統計的に有意であった(*p* < .05)。しかし，2つの要因がどちらも有意になった選択肢はなかった。そのため，より予測率の高いモデルを得るために，各選択肢の選択率を従属変数，有意であった要因を独立変数として強制投入法による単回帰分析を行った。その結果，語彙サイズテストの成績はMorCon, Mor, Non選択肢の選択率との回帰モデルがそれぞれ有意であり，$F(1, 106) = 31.13$, $p < .001$; $F(1, 106) = 26.45$, $p < .001$; $F(1, 106) = 10.60$, $p = .002$，読解熟達度テストの成績はCon選択肢との回帰モデルが有意であった，$F(1, 106) = 6.642$, $p = .011$(表8-11)。しかし，ConとNon選択肢に対するモデルの決定係数(R^2)はどちらも微々たるものであったため(Con: $R^2 = .05$; Non: $R^2 = .08$)，これ以降は，MorCon選択肢およびMor選択肢に焦点を当てて考察を行った。

■表8: 強制投入法による回帰分析の結果 (MorCon選択肢)

独立変数	*B*	95% CI	*SE B*	β	*t*	*p*
(定数)	.193	[.107, .279]	0.043		4.456***	< .001
語彙サイズテスト	.000	[.000, .000]	.000	0.476	5.580***	< .001

$R^2 = .22$ ($N = 108$, $p < .001$), CI = confidence interval for *B*, ***$p < .001$.

■表9: 強制投入法による回帰分析の結果 (Mor選択肢)

独立変数	*B*	95% CI	*SE B*	β	*t*	*p*
(定数)	.357	[.286, .427]	.036		10.024***	< .001
語彙サイズテスト	-.000	[.000, .000]	0.000	-0.447	-5.143***	< .001

$R^2 = .19$ ($N = 108$, $p < .001$), CI = confidence interval for *B*, ***$p < .001$.

■表10: 強制投入法による回帰分析の結果 (Non選択肢)

独立変数	*B*	95% CI	*SE B*	β	*t*	*p*
(定数)	.247	[.107, .279]	0.032		4.456***	< .001
語彙サイズテスト	-.000	[.000, .000]	0.000	-0.301	-3.255**	.002

$R^2 = .08$ ($N = 108$, $p = .002$), CI = confidence interval for *B*, **$p < .01$, ***$p < .001$.

■表11: 強制投入法による回帰分析の結果 (Con選択肢)

独立変数	*B*	95% CI	*SE B*	β	*t*	*p*
(定数)	0.202	[.166, .238]	0.018		4.456***	< .001
語彙サイズテスト	0.009	[.002, .016]	0.003	0.476	5.580*	.011

$R^2 = .05$ ($N = 108$, $p = .011$), CI = confidence interval for *B*, *$p < .05$, ***$p < .001$.

MorCon選択肢の選択率について，決定係数（R^2）は22%であった。つまり，語彙サイズテストの成績がMorCon選択肢の選択率の22%を予測する結果となった。この結果から，語彙サイズテストの成績が高い協力者ほど，MorCon選択肢を選ぶ率が有意に高かったことが示された。この決定係数の値は決して大きいものではないが，MorCon選択肢が今回の未知語推測テストにおいて正答であることを考慮すると，学習者の語彙知識の広さが未知語推測における形態素と文脈情報を利用した推測の成功に対し，少なからず影響を与えていることが示唆された。

Mor選択肢の選択率について，決定係数（R^2）は22%であった。つまり，語彙サイズテストの成績がMor選択肢の選択率の19%を予測する結果となった。ただし，MorCon選択肢と違い，βの値が負であった。このことから，語彙サイズテストの成績が高い協力者ほど，Mor選択肢を選ぶ率が低かったことが明らかとなった。

これらの回帰分析の結果から，協力者の語彙サイズテストの成績は，多肢選択式未知語推測テストの選択肢の選択率をある程度予測することが示された（RQ2）。中国人EFL学習者の未知語推測を調査したZhang and Koda（2012）は，単語を形態素へ分割する能力が未知語推測の能力を介して，（a）英語学習者の語彙知識に寄与すること，さらに（b）語彙知識を介して読解力へ貢献することが示された。本研究は，Zhang and Kodaの結果を一部支持するものであった。本研究では協力者の語彙サイズテストの成績が形態素の意味を含み文脈にも合致する選択肢（i.e.，MorCon選択肢）および目標語に含まれる接頭辞意味を反映した選択肢（i.e.，Mor選択肢）の選択率をある程度予測することが示された。このことから，英語学習者の未知語推測の能力と語彙サイズの間に関係があることが示唆される。

また，以上の回帰分析の結果と未知語推測テストの各選択肢の選択率（表6）から次のことが示唆される：MorCon選択肢においては，接頭辞が利用可能かつ文脈の有益性が低い条件において，他の条件に比べて選ばれる率が低い。このことは，接頭辞が利用可能な条件であっても，文脈の情報が推測に有益でない場合には学習者は形態素（接頭辞）の意味が反映されているかつ文脈にも当てはまる選択肢を選ばないということがわかった。この結果とは逆に，Mor選択肢に関しては，接頭辞が利用可能な条件であっても，文脈の情報が推測に有益でない場合に選択率が高かったことが示された。これらの結果をまとめると，形態素情報（接辞の意味）が利用可能な場合でも，文脈の情報が推測に貢献する場合でなければ正答が選ばれないといえる。また，語彙サイズテストの成績が高い協力者ほどMorCon選択肢を選び，逆にMor選択肢を選ばないことから，語彙サイズの大きな協力者は接頭辞の情報のみを利用するのではなく，文脈情報と照らし合わせて選択肢を選んだことが示唆される。これは，学習者の英語熟達度によって，未知語推測に用いる情報が異なることを報告したHamada（2014）と一致する。Hamadaの研究では，英語熟達度の低い学習者は，形態素の情報が未知語の意味推測に寄与しない状況においても，形態素の意味にこだわり未知語の意味推測に失敗するのに対し，熟達度の高い学習者は柔軟に文脈の情報を利用し推測に成功する傾向を報告した。この結果と本調査の結果から，文脈の情報の柔軟な利用に語彙知識の広さが寄与したことが示唆される。したがって，未知語推測の処理プロセスの初期の段階である利用可能な手がかりの利用において，形態素情報に依存した推測をしがちな学習者は，自身の英語熟達度（特に語彙サイズ）の低さゆえに文脈情報を利用することが難しく，推測の成功にたどり着くことができないと考えられる。

5 結論と今後の課題

本研究では，未知語推測中に利用される情報に焦点を当て，解答パターンから学習者の従事する未知語推測プロセスを判断することができる多肢選択式テストの作成について示唆を得ることを目的に調査を行った。本研究の知見をまとめると次のようになる：（a）英語学習者は，目標語（目標文）の語形や文脈の手がかりを組み合わせるプロセスを経て，意味を推測しようとする，（b）英語学習者の語彙サイズテストの成績は，多肢選択式未知語推測テストの選択肢の選択率をある

程度予測する，(c)未知語推測を行う際，形態素情報に依存した推測をしがちな学習者は，自身の英語熟達度（特に語彙サイズ）の低さゆえに文脈情報を利用することが難しく，推測の成功にたどり着くことができない。

本研究の限界点をまとめると，次の通りである：(a)本研究の協力者にとって英文読解熟達度の難易度が高すぎた。 本研究では，英検2級が高等学校卒業レベルであることと（日本英語検定協会，2018），本調査の協力者が大学生であることを考慮し，2級と準1級の問題を用いることが英文読解熟達度の弁別に妥当であると考え，英検2級と準1級の過去問をもとに英文読解熟達度を測定するテストを作成したが，結果として低い成績のものが多かった。今後は，より幅広い級（英検準2級-1級）から問題を選び，テストを作成することが考えられる。(b)推測中に着目した手がかりが明らかにできなかった。今回の調査は紙面の未知語推測テストを用いて実施した。そのため，実際にどのような手がかりを協力者が用いたのかは明らかにできなかった。この点を解決するために思考発話法を用いた研究（e.g., Nassaji, 2003）を実施する必要があると思われる。

本研究の示唆をまとめると次の通りになる：本研究の結果から，未知語推測を行う際，形態素情報に依存した推測をしがちな学習者は，自身の英語熟達度（特に語彙サイズ）の低さゆえに文脈情報を利用することが難しく，推測の成功にたどり着くことができないことが明らかとなった。ここから，多肢選択式未知語推測テストを作成する際には，1つの情報に偏った選択肢のみを提示することは避けるべきであるという示唆が得られた。また，この結果から，語彙サイズの低い学習者には，文脈の情報へ注意を向け，利用を促す指導が必要と考えられるという教育的示唆が得られた。

謝辞

本研究の実施，発表にあたって公益財団法人日本英語検定協会と関係者の皆さま，ならびに選考委員の先生方からのご支援を頂きました。厚く御礼申し上げます。特に，助言者である村木英治先生は，研究の実施および本報告書の執筆にあたりご指導をいただき大変感謝しております。そして，筑波大学の卯城祐司先生，東京都市大学の細田雅也先生，白鴎大学の森好紳先生には，本研究の調査実施に関して多くのアドバイスを頂きました。また，同じ研究室でともに学んできた，鈴木健太郎さん，多田豪さん，天海有加さん，大河原にじ香さんをはじめとするメンバーの皆さまにも多くの支援を頂きました。最後に，本研究の実験に参加してくださった協力者の皆さまに深く御礼申し上げます。

参考文献（＊は引用文献）

＊ 相澤一美・望月正道（編著）(2010).『英語語彙指導の実践アイディア集：活用例からテスト作成まで』東京：大修館書店.
＊ 旺文社 (2015a).『2015年度版 英検2級過去6回全問題集』. 東京：旺文社.
＊ 旺文社 (2015b).『2015年度版 英検準1級過去6回全問題集』. 東京：旺文社.
＊ 旺文社 (2017a).『2017年度版 英検2級過去6回全問題集』. 東京：旺文社.
＊ 旺文社 (2017b).『2017年度版 英検準1級過去6回全問題集』. 東京：旺文社.
＊ 竹内理・水本篤（編著）(2014).『外国語教育ハンドブック【改訂版】』. 東京：松柏社.
＊ 日本英語検定協会 (2018).「2級の試験内容・過去問」. (http://www.eiken.or.jp/eiken/exam/grade_2/). 2018.5.9 取得.
＊ 望月正道 (1998).「日本人学習者のための英語語彙サイズテスト」.『語学教育研究所紀要』, 12 号, 27-53.
＊ 文部科学省 (2009).『高等学校学習指導要領』. 京都：東山書房.
＊ Aebersold, J. A., & Field, M. L. (1997). *From reader to reading teacher: Issues and strategies for second language classroom.* Melbourne: Cambridge University Press.
＊ Cho, K-S., & Krashen, S. (1994). Acquisition of vocabulary from the Sweet Valley Kids Series: Adult ESL acquisition. *Journal of Reading, 37,* 662-667. Retrieved from http://www.jstor.org/stable/20172388
＊ de Bot, K., Paribakht, T. S., & Wesche, B. M. (1997). Toward a lexical processing model for the study of second language acquisition: Evidence from ESL reading. *Studies in Second Language Acquisition, 19,* 309-329. doi:10.1017/s0272263197003021
＊ Ender, A. (2014). Implicit and explicit cognitive processes in incidental vocabulary acquisition. *Applied Linguistics, 37,* 536-560. doi:10.1093/applin/amu051
＊ Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge University Press.
＊ Haastrup, K. (1991). *Lexical inferencing procedures or talking about words: Receptive procedures in foreign language learning with special reference to English.* Tubingen, Germany: Gunter Narr.
＊ Hamada, M. (2014). The role of morphological and contextual information in L2 lexical inference. *The Modern Language Journal, 98,* 992-1005. doi:10.1111/modl.12151
＊ Huckin, T., & Bloch, J. (1993). Strategies for inferring word meaning in context. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 153-178). Norwood, NJ: Ablex.
＊ Kamimura, K. (2017). Effects of prefix availability and vocabulary size on Japanese EFL lexical inferencing. *ARELE, 28,* 273-288. doi:10.20581/arele.28.0_273
＊ Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal, 78,* 285-299. doi:10.1111/j.1540-4781.1994.tb02043.x
＊ Levelt, W. J. M. (1993). Language use in normal speakers and its disorders. In G. Blanken, J. Dittman, H. Grimm, J. Marshall, & C. Wallesch (Eds.), *Linguistic disorders and pathologies: An international hand book* (pp. 1-15). Berlin, Germany: de Gruyter.
＊ McCutchen, D., & Logan, B. (2011). Inside incidental word learning: Children's strategic use of morphological information to infer word meanings. *Reading Research Quarterly, 46,* 334-349. doi:10.1002/RRQ.003
＊ Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System, 28,* 291-304. doi:10.1016/S0346-251X(00)00013-0
＊ Nakagawa, C. (2006). The effects of morphological and contextual clues on EFL readers' lexical inference. *ARELE, 17,* 151-160. doi:10.20581/arele.17.0_151
＊ Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly, 37,* 645-670. doi:10.2307/3588216
＊ Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63,* 59-82. doi:10.3138/cmlr.63.1.59
＊ Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
＊ Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language* Acquisition, 21, 195-224. doi:10.1017/S027226319900203X
＊ Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12,* 329-363. doi:10.1177/1362168808089921
＊ Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Palgrave Macmillan: London, UK.
＊ Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20,* 232-245. Retrieved from https://search.proquest.com/docview/1705668775?accountid=25225
＊ Zhang, D., & Koda, K. (2012). Contribution of morphological awareness and lexical inferencing ability to L2 vocabulary knowledge and reading comprehension among advanced EFL learners: Testing direct and indirect effects. *Reading and Writing, 25,* 1195-1216. doi:10.1007/s11145-011-9313-z

資料1　語彙サイズテスト（相澤・望月，2010）の例

日本語の意味を表す英語を(1)～(6)の中から選び、その番号を解答欄に書いてください					
		1. だいだい色の果物		2. トウモロコシ	
(1) chicken	(2) corn	(3) orange	(4) plant	(5) tomato	(6) vegetable
		3. 桃		4. 昼食	
(1) bag	(2) bowl	(3) box	(4) lunch	(5) peach	(6) sandwich
		5. 野球		6. 札、券	
(1) baseball	(2) calendar	(3) card	(4) chalk	(5) data	(6) snow

資料2　目標語と目標語が含む接頭辞

接頭辞	目標語	接頭辞	目標語
anti-	*antihero*	*arch-*	*archenemy*
re-	*redescription*	*ab-*	*abnormality*
non-	*nonmember*	*per-*	*peruse*
un-	*unrest*	*ante-*	*anteroom*
sub-	*subeditor*	*bi-*	*biplane*
mid-	*midship*	*de-*	*depopulation*
mis-	*misdirection*	*circum-*	*circumnavigation*
pre-	*prehistory*	*ob-*	*oblong*
neo-	*neogeography*	*pro-*	*protractor*
dis-	*discomfort*	*ex-*	*exposition*
ad-	*adjudgement*	*com-*	*compassion*

資料3　未知語推測テストの例

This biplane was designed by a Spanish engineer. He was very glad that it soon became popular.

① この単語を知っていましたか?	□はい　□いいえ
② この英文中の下線が引かれた語の意味として適切なものを次の1～4の中から1つ選び，○で囲ってください。	
1. an electronic machine that can store and work with a lot of information 2. a type of airplane that has two sets of wings 3. a large animal that is used for riding and for carrying things 4. a thing that has two legs and uses them to walk	

資料4　接頭辞テストの例

biscowd　bikuncks　bisprolled

下線を引いた部分の意味を下の欄に日本語で記入してください。
意味：

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

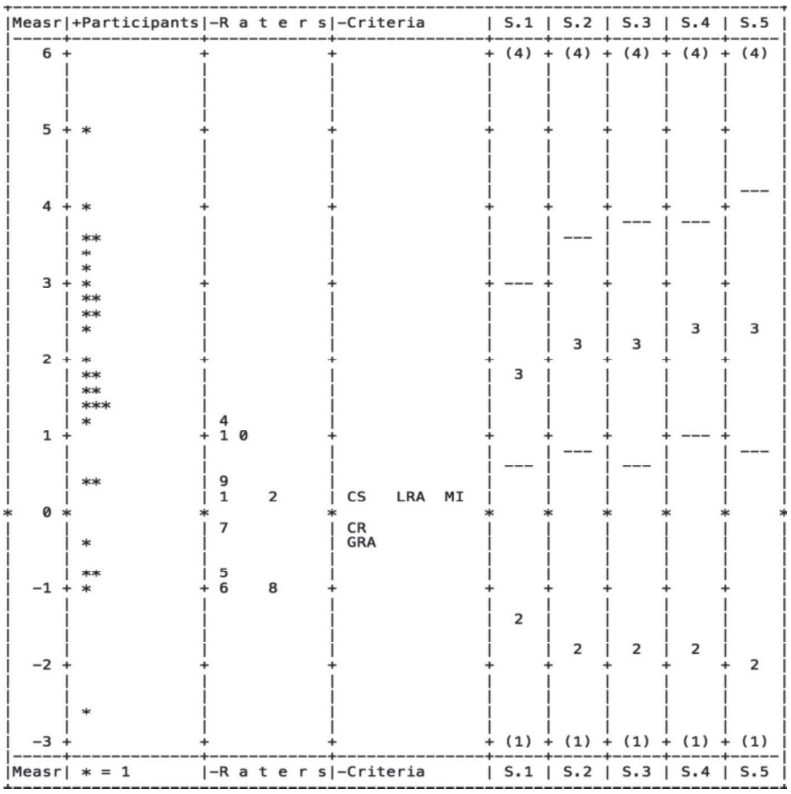
Knoch（2009）とHirai and Koizumi（2013）の多相ラッシュ分析の解釈方法を参考に，下記の観点で2つのルーブリックを比較する（表7）。

5 結果と考察

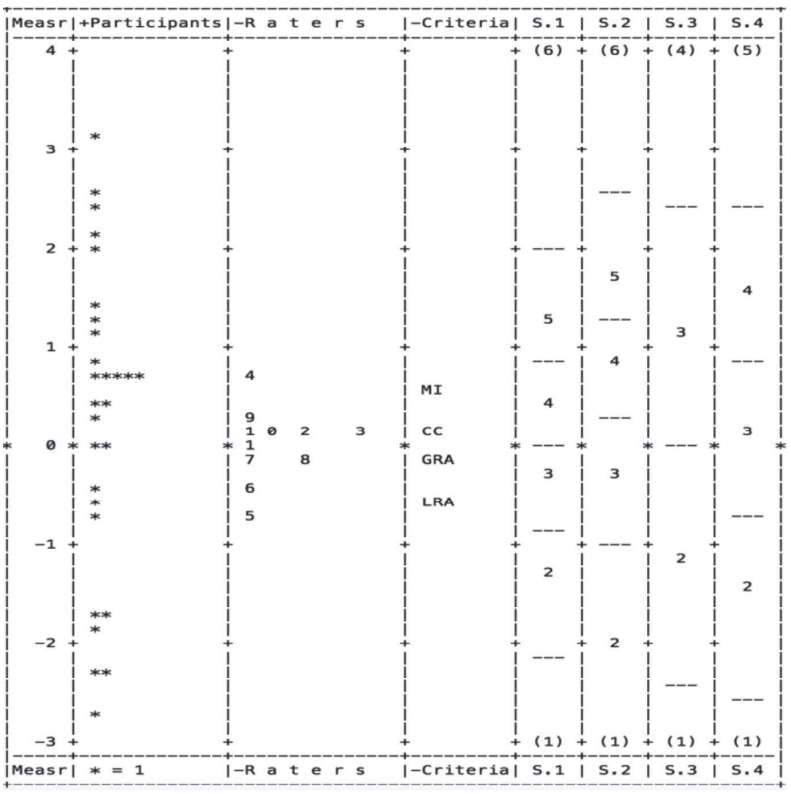
各変数マップ（図1, 図2）は，左から，ロジット（logit）という共通尺度で表した推定値（measure），テスト参加者，採点者，評価観点の3相を示している。推定値が上がる（下がる）に

■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

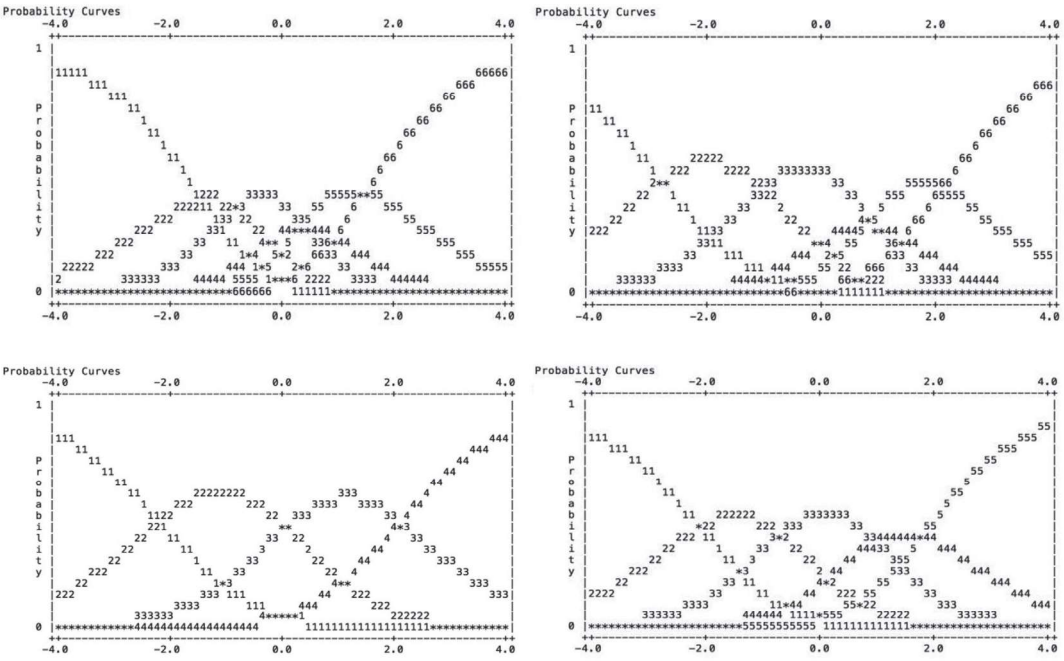
＊Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

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。

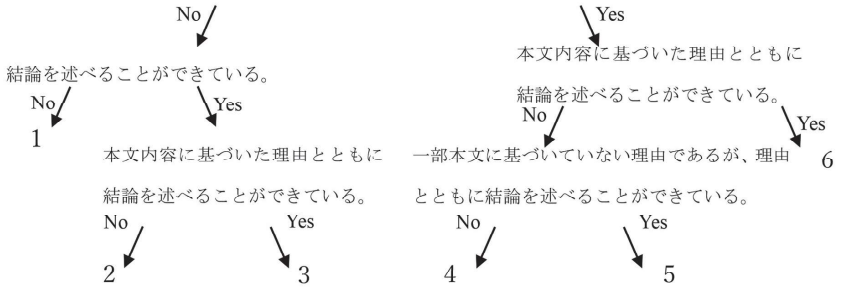
資料3　採点者アンケート結果のまとめ(一部省略)

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。

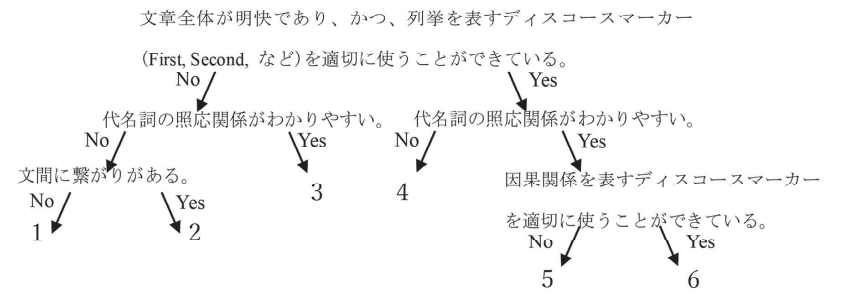
資料4　本研究で用いたEBBルーブリック

The EBB rating scale for TEAP Writing Task B

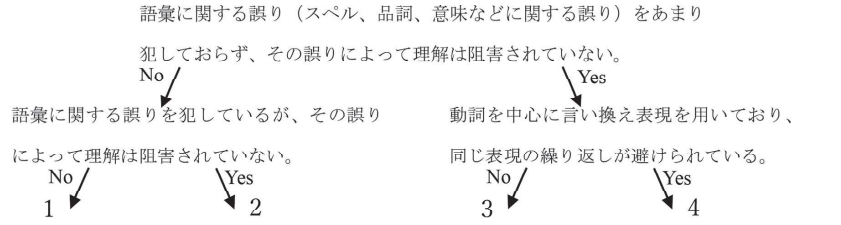
1. MAIN IDEAS



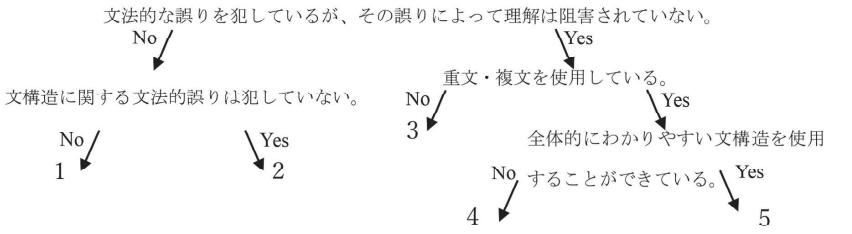
2. COHERENCE & COHESION



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。

第30回 研究助成
A 研究部門 ● 報告Ⅳ ● 英語能力テストに関する研究

スピーキング練習における特定文法項目に焦点を当てた直接的修正フィードバックの効果検証

研究者:愛知県／名古屋大学大学院　在籍　小林 真実
《研究助言者:池田 央》

概要

本研究の目的は、英語のスピーキング指導における焦点化修正フィードバックの効果検証である。教師が学習者に多くのフィードバックを与えた場合，アップテイクにつながらないとの指摘がある（Truscott,1996, 1999）。本研究は，動詞の過去形に焦点を当てた焦点化修正フィードバックと，動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態のエラー，受動態のエラーについて修正フィードバックを行う非焦点化修正フィードバックの効果を比較し、学習者の文法的正確性に効果的な修正フィードバックを検証した。実験参加者を3群に分けて，焦点化修正フィードバックを与える実験群，非焦点化修正フィードバックを与える比較群，修正フィードバックを与えない統制群を設けた。その結果，事前テストと3，4週間後の遅延事後テストの比較において，動詞過去形のエラーの減少は有意傾向がみられたが，3群の差はみられなかった。

1 はじめに

1.1 研究目的

本研究の目的は，英語のスピーキング練習においてエラーに対し修正を促す修正フィードバック（corrective feedback）の対象を1つの文法項目に限定した焦点化修正フィードバックの効果検証である。スピーキング練習において，1つの文法項目について教師がエラーを指摘し，正しい形を示す焦点化直接的修正フィードバックを行った場合と，複数の文法項目について直接的修正フィードバックを行う非焦点化直接的修正フィードバックを行った場合の，直後事後テストおよび遅延事後テストにおける発話の文法的正確性の向上を比較し，効果的な修正フィードバックの方法を検証した。

1.2 フィードバック

第二言語習得において，フィードバックとは学習者のアウトプットについて教師などが言及するものであり，その方法は多岐に渡る。学習者の正しい答えを褒めることにより学習者の答えが正しい，もしくは良いと伝える方法や，学習者が間違ったアウトプットをしたときに，エラーが発生したことを知らせる明確化の要求，エラーに関する明示的説明，学習者自身がエラーに気付くことを促すリキャスト，自己訂正の促し，学習者の犯したエラーの繰り返し，メタ言語フィードバックが挙げられる。また，フィードバックは教師だけが行うとは限らず，他の学習者がフィードバックを行う場合もある。また，英語の指導現場において，教師は学習者に繰り返しフィードバックを与えるが，1度のみフィードバックを与えた場合も長期的効果があると報告されている（Chandler, 2003; Van Beuningen, De Jong, & Kuiken, 2012など）。また，フィードバックの種

類により異なった効果があると報告されている（Saeb, 2014; Shintani, Ellis, & Suzuki (2014); Van Beuningen, De Jong, & Kuiken, 2012 など）。

1.2.1 直接的修正フィードバックと間接的修正フィードバック

修正フィードバックとは，教師がエラーを指摘する種類のフィードバックであり，その方法には2種類ある。直接的修正フィードバックとは，教師が学習者のエラーを指摘し，修正するものである。例えば，学習者が“I go to the coffee shop yesterday”と言った場合，教師は“go”はエラーであることを指摘し，“I went to the coffee shop yesterday”が正しい文であることを伝える。一方で，間接的修正フィードバックとは，教師は学習者のエラーを指摘するが，エラーを正しい形に修正しないものである。上に挙げた例の場合，教師は“go”はエラーであることを指摘するが，正しい形は“went”であることを提示せず，学習者自身が正しい形を考えることを促す。学習者のライティングにおける修正フィードバックの効果については研究が行われており，直接的修正フィードバックは間接的修正フィードバックより文法的正確性が向上すると報告されている。Van Beuningen et al.（2012）は，ライティング指導における直接的および間接的修正フィードバックの文法的正確性への効果の検証を行った。直接的および間接的修正フィードバックを与えられた実験参加者は，ライティングを学習者自身が修正する自己修正を行った参加者や，フィードバックと自己修正を行わずにライティングの練習を多く行った参加者と比して，新しいライティング課題を行ったときの文法的および語彙的正確性の向上への効果が高かった。また，直接的修正フィードバックを与えられた実験参加者は，文法的正確性が向上したのに対し，間接的修正フィードバックを与えられた実験参加者は，語彙的正確性が向上した。この研究結果から，直接的修正フィードバックは間接的修正フィードバックと比して文法的正確性により多くの注意が向けられると言えるであろう。

1.2.2 直接的修正フィードバックとメタ言語的説明

メタ言語的説明とは，教室内でみられる学習者のエラーについて教師が説明を行うものである。Shintani, Ellis, & Suzuki（2014）は，ライティングにおける直接的修正フィードバックとメタ言語的説明の効果を比較するため，学習者のライティングの仮定法と不定冠詞のエラーについて直接的修正フィードバックとメタ言語的説明を行い，比較検証した。その結果，仮定法のエラーには直接的修正フィードバックとメタ言語的説明はどちらも効果が認められたが，不定冠詞にはどちらのフィードバックも効果が確認されなかった。また，直接的修正フィードバックはメタ言語的説明と比して長期的効果が認められた。この結果から，文法項目によりフィードバックの効果は異なるが，直接的修正フィードバックはメタ言語的説明と比してより多くの長期的効果があることが明らかになった。

1.2.3 焦点化修正フィードバックと非焦点化修正フィードバック

教師によるフィードバックは英語教育現場において広く使用されているが，修正フィードバックの問題点として，複数の修正フィードバックを学習者に与えた場合，学習者がフィードバックを認識して理解するというアップテイクが起こらないという指摘がある（Truscott, 1996, 1999）。その理由として，教師から多くのフィードバックを与えられると，学習者は全ての修正フィードバックの内容に注意を向けるのが困難となり，その結果フィードバックから自分の発話をモニタリングして必要に応じて修正することができなくなる。そのため，文法への修正フィードバックは効果がないと説明している。

この問題の解決策として，一度に多くのフィードバックを与えず，限られた文法項目についてフィードバックを与えるという方法があり，焦点化修正フィードバックと呼ばれる（Sheen, 2007）。特定の文法項目について修正フィードバックを与える方法であり，対象文法項目は1項目の場合や数項目の場合もある。一方で，非焦点化修正フィードバックとは，焦点化修正フィードバックより多い項目数，もしくは全ての文法項目につ

いて修正フィードバックを与えるものであり，事前に複数の文法項目を選択してフィードバックを与える場合と，文法項目を特定せずに全ての文法的エラーについて修正フィードバックを与える場合がある。

第二言語のライティングにおける焦点化修正フィードバックは，焦点化修正フィードバックの効果が認められており，学習者はライティングについて焦点化修正フィードバックを与えられた後に新しいライティングタスクを行う場合，非焦点化修正フィードバックを受けた場合と比して文法的正確性が向上すると報告されている。Sheen, Wright, & Moldawa (2009) は日本の大学生を対象に，冠詞の修正フィードバックを与える焦点化群，冠詞，動詞の過去形，前置詞について修正フィードバックを与える非焦点化群，修正フィードバックを与えない統制群を設けてライティングタスクを与えた。その結果，焦点化群と非焦点化群は新しいライティングタスクにおいて，統制群と比して文法的正確性が向上した。直後事後テストでは焦点化群と非焦点化群の間に差はみられなかったが，4週間後に行われた遅延事後テストでは焦点化群は非焦点化群より文法的正確性が高かった。その一方で，事後テストでは焦点化群と非焦点化群では差がみられないと報告する研究もある (Saeb, 2014)。

1.2.4 直接的修正フィードバックの対象文法項目

学習者がエラーに気付くことにより修正フィードバックの効果が表れると指摘されている (Ellis, 2005など)。また，修正フィードバックは，対象文法項目により英語学習者がエラーに気付く割合が異なるとされており，動詞の過去形のエラーに修正フィードバックを与えた場合は，疑問文に関するエラーと比して，学習者は気付きやすいと報告されている (Kartchava & Ammar, 2014)。直接的修正フィードバックとメタ言語的説明の効果を比較検証したShintani, Ellis, & Suzuki (2014) では，対象文法項目により，2つのフィードバック方法の効果が異なっていた。仮定法のエラーにはどちらのフィードバック方法も効果がみられた一方で，不定冠詞のエラーについてはどちらの方法も効果が認められなかった。

明示的に誤りを訂正する修正フィードバックが効果的であるのは，「規則の内部構造が単純な項目」，「語彙的意味の伝達が主となる項目」，「日本語 (母語) に同じか類似した概念・構造が存在する項目」，「今までに十分教えられてこなかった項目」であるとされている。反対に，「規則の内部構造が複雑な項目」，「文法的機能の伝達が主となる項目」，「日本語 (母語) に同じか類似した概念・構造が存在しない項目」，「その規則についてすでに十分な知識を持っている項目」は明示的な誤り訂正が効果的ではないとされている (白畑, 2015, p. 182)。

修正フィードバックを行う文法項目により効果が異なり，またフィードバックを与えてもその効果がみられない文法項目もあるため，フィードバックに関する検証を行う際には，フィードバックが効果を持つ対象文法項目を選択し，文法項目による影響が結果を左右しないようにする必要がある。

1.2.5 第二言語スピーキングにおける修正フィードバックの研究

スピーキングにおける紙面上の修正フィードバックに関しては，ライティングと比べて十分に研究が行われていないが，Date & Takatsuka (2012) は，スピーキングにおける自己修正と間接的修正フィードバックの効果を検証した。漫画描写タスクの発話について，実験参加者が自ら書き起こして修正した後，修正が不十分である参加者には実験実施者が間接的修正フィードバックを与え，参加者は修正されたスクリプトを黙読した。その後，同じ漫画の描写タスクおよび新しい漫画の描写タスクを実施し，最初のタスクの発話と文法的正確性を比較したところ，同じタスクと新しいタスクの発話において，書き起こし，自己修正，間接的修正フィードバック，黙読の機会を与えられなかった参加者と比して，文法的正確性が向上した。この結果から，修正フィードバックはスピーキングにおいても効果があると予測されるが，この研究において，1度目の発話後に書き起こし，自己修正，黙読の要因も行っているため，文法的正確性の向上は，修正フィードバックによるものではない可能性がある。また，間接的修正フィードバックを行ったため，

直接的修正フィードバックの効果も明らかにされていない。

1.2.6 先行研究の改善点およびそれらの解決案

第二言語ライティングに関する研究において，焦点化直接的修正フィードバックを行う場合，非焦点化直接的修正フィードバックを行う場合と比べて，新しいタスクにおける文法的正確性が向上する可能性が報告されている。しかし，第二言語のスピーキングについては，焦点化直接的修正フィードバックの効果が十分に検証されておらず，スピーキング指導における効果的な修正フィードバックの方法は明らかにされていない。しかし，ライティングに関する結果から，スピーキング指導においても，焦点化直接的修正フィードバックの効果があると予想される。

1.3 研究課題

上述の先行研究の知見を踏まえて，本研究が設定する研究課題は以下のとおりである。

1. 英語スピーキングにおいて，動詞の過去形のエラーを修正フィードバックの対象文法項目とした焦点化修正フィードバックは，動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラーを修正フィードバックの対象文法項目とした非焦点化修正フィードバックと比して，直後に別のタスクを行ったときに，動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラー率は減少するか。
2. 英語スピーキングにおいて，動詞の過去形のエラーを修正フィードバックの対象文法項目とした焦点化修正フィードバックは，動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラーを修正フィードバックの対象文法項目とした非焦点化修正フィードバックと比して，3週間から4週間後に別のタスクを行ったときに，動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラー率は減少するか。

研究課題1と2に取り組むために，本研究では，焦点化直接的修正フィードバックを与える群，非焦点化直接的修正フィードバックを与える群，直接的修正フィードバックを与えない群の3群に分けて，(1)事前テスト，(2)フィードバック，(3)フィードバック後のテスト，(4)直後事後テスト，(5)遅延事後テストを行い，(1)事前テストのエラー率を(4)直後事後テストのエラー率および(5)遅延事後テストのエラー率と比較し，(4)直後事後テスト，(5)遅延事後テストにおける文法的正確性の向上を調査した。

2 本研究の内容

2.1 実験参加者

本研究の実験参加者は，日本語を母語とする大学学部生であり，年齢幅は18歳から22歳（最頻値は19）であった。参加者は，少なくとも6年間は日本での教育機関で英語を学習しており，大学での専攻は，工学，農学，文学などさまざまであった。調査時の英語力（自己申請）は，TOEFL ITP® 平均499.53（*SD* =46.17）であった。4名の実験参加者はTOEFL ITP®の受験経験がなかったが，TOEIC® の点数は平均633.75（*Min* =540, *Max* =690, *SD* =56.499）であった。

実験参加者は，実験内容，所要時間，謝礼および個人情報保護に関しての説明を受け，同意した後に約90分間の実験を行った。実験の参加に同意した51名を無作為に3群に分けて，動詞過去形について修正フィードバックを与える実験群（*n* =17, TOEFL ITP® *M* =451.611, *SD* =167.917），動詞の時制，動詞の欠如，目的節の欠如，前置詞について修正フィードバックを与える比較群（*n* =16, TOEFL ITP® *M* =378.944, *SD* =204.911），修正フィードバックを与えない統制群（*n* =15, TOEFL ITP® *M* =470.563, *SD* =128.163）を設けた。実験参加者51名の内，2名は2度目の実験実施に参加しなかったため，分析から除外した。また，1名は発話の録音に失敗したため，分析から除外した。残りの48名の発話を分析の対象とした。

2.2 本研究で使用したタスク

本実験は，スピーキングタスクとしてナラティブタスクを行った。6コマ漫画（Heaton, 1975）を使用した。漫画を見たことのない人がストーリーを理解できるように説明するよう指示した。実験参加者は，60秒間で漫画を見てストーリーを把握し，実験実施者である筆者は，実験参加者がストーリーを理解したことを確認し，参加者がストーリーを理解していなかった場合のみ日本語でストーリーを説明した。その後，参加者は90秒間でストーリーを説明した。漫画を見ながら発話したため，発話はストーリーを記憶する必要のない状態で行った。

2.3 直接的修正フィードバック

本研究では，エラーを指摘し正しい文法形式を示す直接的修正フィードバックを与えた。外国語のライティングに関する研究において，直接的修正フィードバックは，エラーの指摘のみを行う間接的修正フィードバックと比して正確性の向上を報告する研究が多い（Van Beuningen, et al., 2012など）ことが理由である。

また，フィードバックの実施方法は，口頭ではなく筆記で行った。スピーキング指導において，教師は口頭でフィードバックを提供することが多いが，フィードバックの内容を実験に合うように統制するには口頭では困難であり，筆記によるフィードバックはより正確に実施ができると考えられるため，筆記で行った。

焦点化修正フィードバックは，対象文法項目は1項目の場合や数項目の場合もあるが，本研究では1項目について修正フィードバックを与えるものを焦点化修正フィードバックとした。また，非焦点化修正フィードバックは，事前に複数の文法項目を選択してフィードバックを与える場合と，文法項目を特定せずに全ての文法的エラーについて修正フィードバックを与える場合があるが，本研究では下に挙げる6項目について修正フィードバックを与えるものとした。

修正フィードバックは，1.2.4で修正フィードバックが効果的であるとされていると述べたように，対象とする文法項目によりその効果が異なるとされている。そのため本研究では，「規則の内部構造が単純である項目」，「語彙的意味の伝達が主となる項目」，「母語に同じか類似した概念・構造が存在する項目」である文法項目を選び，実験群および比較群の修正フィードバックの対象項目を以下のとおりとした。

実験群　動詞の過去形
比較群　動詞の過去形，定動詞の欠落，
目的語の欠落，前置詞の欠落，
能動態および受動態のエラー

本研究において，動詞の過去形のエラーとは，動詞が過去形であるべき箇所において過去形になっていないエラーを指すが，動詞の屈折のエラーは扱わなかった。例えば，“go”が“goed”となっていても，語彙的なエラーであり，動詞を過去形にする意思は見受けられるため文法的なエラーではないと考えて，エラーとしてみなさなかった。

定動詞の欠落とは，定動詞を使用しなければならない箇所（義務的文脈）において定動詞を使用していないものを指す。ただし，定動詞を使用しなければならない箇所に不定動詞がある場合は，語彙的なエラーであるのか文法的なエラーであるのか判別ができないため，定動詞の欠落としてみなさない。例えば，“he gone to the sea yesterday”という節において，動詞は正しくは“went”であるが，実験参加者が“go”の過去形として“gone”と言ったのか（語彙的エラー），現在完了形“has gone”を誤って“gone”と言ったのか（文法的エラー）判断できないため，定動詞欠落のエラーとしてカウントしなかった。

目的語の欠落とは，目的語を使用しなければならない箇所（義務的文脈）において目的語が欠落しているものを指す。例えば，学習者が“he had in his hand”を産出した場合，“had”は他動詞であり目的語を使用しなければならないが，目的語を使用していないため，目的語の欠落とみなす。目的語を必要としない文において誤って目的語を入れた場合はエラーとしてみなさなかった。例えば，少年が泳いでいる絵を描写するとき“he swam himself”と言った場合，この文脈において“swam”は目的語を使用しないが，エラーとして扱わなかった。

前置詞の欠落とは，前置詞を使用しなければならない箇所（義務的文脈）において前置詞が欠落しているものを指す。例えば，“he went the sea”という節では，“sea”の前に”to”という前置詞を使用しなければならないが，前置詞を使用していないため，前置詞の欠落とみなした。前置詞を必要としない文において誤って前置詞を入れた場合は前置詞のエラーとして扱わなかった。例えば，“the bus splashed to water”という節において“water”は目的語であるためその前に“to”は使用すべきではないが，前置詞の欠落ではないため，エラーとしてカウントしなかった。また，前置詞が必要な文において不適切な前置詞を使用した場合，語彙的なエラーであるか文法的なエラーであるのか判別ができないため，前置詞の欠落としてみなさなかった。 例えば，“there were many people on the town”という節において“on”という前置詞は不適切であるが，前置詞の欠落ではないため，エラーとして扱わなかった。

能動態のエラーとは，能動態を使用しなければならない箇所（義務的文脈）において受動態を使用した場合を指す。また，受動態のエラーとは，受動態を使用しなければならない箇所（義務的文脈）において能動態を使用した場合を指す。例えば，「男が少年をつかまえた」という意味の節において，“the boy was caught by the man”が正しい節であり，“the man was caught by the boy”は能動態のエラー，“the boy caught the man”は受動態のエラーである。また，受動態を使用しているが動詞の屈折におけるエラーがある場合は，語彙的エラーとみなし，受動態のエラーとしてみなさなかった。例えば，“the boy was catched by the man”は“caught”を誤って“catched”とした語彙的エラーであり，受動態のエラーとしてカウントしなかった。

2.4 実験手順

実験前に，実験群と比較群の実験参加者に，1つ目の発話について簡単なコメントをすると告げた。統制群の参加者には，発話を書き起こしたものを見せると伝えた。また，コメントおよび書き起こしたスクリプトを見せた後に，再度同じタスクを行うことも伝えた。全ての実験参加者は事前テストとして，1つ目の漫画描写タスク（タスクA）を行った。その後，参加者は5分間休憩を取った。休憩時間中は，辞書で英単語を調べること，および友人に電子メール等で英語に関する

■表1: 実験手順

	実験群	比較群	統制群
第１回	事前テスト（タスクＡ１度目）		
	休憩（書き起こし，フィードバック）	休憩（書き起こし，フィードバック）	休憩（書き起こし）
	書き起こしとフィードバックを読む	書き起こしとフィードバックを読む	書き起こしを読む
	フィードバック後テスト（タスクＡ２度目）		
	直後事後テスト（タスクＢ）		
3-4週間の間隔			
第２回	遅延事後テスト（タスクＣ）		

■表2: タスクの実施順序

		サブグループ１	サブグループ２	サブグループ３
第１回	事前テスト	タスクＡ	タスクＢ	タスクＣ
	フィードバック後テスト	タスクＡ	タスクＢ	タスクＣ
	直後事後テスト	タスクＢ	タスクＣ	タスクＡ
第２回	遅延事後テスト	タスクＣ	タスクＡ	タスクＢ

質問をすること以外は何をしても良いと伝えた。休憩時間の間に実験実施者は発話の録音を聞いて，パソコンで書き起こし，実験群には焦点化直接的修正フィードバック，比較群には非焦点化直接的修正フィードバックを与えた。休憩時間後，実験群には書き起こしたスクリプトと焦点化直接的修正フィードバック，比較群には書き起こしたスクリプトと非焦点化直接的修正フィードバック，統制群には書き起こしたスクリプトを見せた。実験群と比較群には，一部のエラーを修正したと伝えた。実験参加者は書き起こしたスクリプトと修正フィードバックを読んだ後，フィードバック後テストとして再度同じタスク（タスクA）を行った。その後，直後事後テストとして2つ目の漫画描写タスク（タスクB）を行った。3-4週間後に遅延事後テストとして3つ目の漫画描写タスク（タスクC）を行った。手順を表1に示す。

また，漫画の描写における難易度の発話への影響を排除するため，3群を更に3つのサブグループに分けて，使用する漫画の順序を変えてカウンターバランスをとった。サブグループ1は，タスクA，A，B，Cの順番でタスクを行い，サブグループ2はタスクB，B，C，A，サブグループ3はタスクC，C，A，Bの順でタスクを実施した（表2）。

3 結果

3.1 データ分析方法

多変量分散分析により3群のエラー率を比較した。従属変数は，エラー率とした。エラー率は，発話においてみられたエラーの数を，その発話において必要とされる箇所の数（義務的文脈）で割ったものとした。

3.2 データ分析結果

焦点化直接的修正フィードバックと非焦点化直接的修正フィードバックの対象である動詞の過去形，および非焦点化直接的修正フィードバックの対象である定動詞の欠落，目的節の欠落，前置詞の欠落，能動態と受動態のエラー数の平均および標準偏差と，これらの文法項目が必要とされる義務的文脈の件数（発生件数）の平均および標準偏差，エラー率を表3から表14に，焦点化直接的修正フィードバックの対象文法項目であった動詞過去形のエラー率を図1に示す。また，多変量分散分析の結果を資料1から10に示す。

事前テストの後に実験群と比較群に与えた直接的修正フィードバックの件数は，実験群は平均3.71回，比較群は平均5.00回であった（表3 実験群の過去形のエラー数および比較群の全エラー数を参照）。前置詞欠落および受動態のエラーについては，使用が必要とされる義務的文脈の発生がみられない発話があり，分析を行うことができなかったため，分析の対象から除外した。表5, 8, 11, 14では，NAと表記する。

事前テストと直後事後テストの比較において，実験群と比較群は全体的にエラー率が減少したが，統制群では差がみられなかった（表5, 11）。多変量分散分析を行った結果，テスト間に有意差があり（資料1），直後事後テストにおいて6つの文法項目への焦点化および非焦点化直接的修正フィードバックの効果があったと考えられる。

焦点化直接的修正フィードバックの対象であった動詞の過去形のエラー率は，実験群と比較群の発話において減少しているが，統制群の発話では差がみられなかった（図1）。しかし，多変量分散分析の結果，実験群と比較群共にテスト間の有意差はみられず，小程度の効果量が確認された（$F(1, 88) = 1.98$, $p < .16$, $\eta^2 = .02$）。また，群間においても有意差がなく，効果量もなかった（$F(2, 88) = .03$, $p < .96$, $\eta^2 = .00$）。

動詞欠落のエラー率は，実験群および統制群において減少がみられたが，比較群においては差がみられなかった。多変量分散分析の結果，テスト間（$F(1, 88) = 2.35$, $p < .12$ $\eta^2 = .02$）と群間（$F(2, 88) = 1.36$, $p < .26$ $\eta^2 = .02$）において有意差がみられず，小程度の効果量があった。目的節欠落のエラーは，実験群と比較群はエラー率が減少したが，統制群は差がみられなかった。多変量分散分析の結果，テスト間は有意傾向であり，小程度の効果量があった（$F(1, 88) = 3.68$, $p < .05$ $\eta^2 = .03$）が，群間においては有意差がなく，効果量もなかった（$F(2, 88) = .46$, $p < .62$ $\eta^2 = .00$）。能動態であるべき箇所を誤って受動

態にするエラーは，比較群と統制群において小程度の減少がみられたが，実験群においては変化がみられなかった。多変量分散分析の結果，テスト間（$F(1, 88) = 1.81$, $p < .18$ $\eta^2 = .02$）および群間（$F(2, 88) = .48$, $p < .61$ $\eta^2 = .01$）において有意差がみられなかったが，小程度の効果量が認められた。

事前テストと遅延事後テストの全体の比較に

■表3: 事前テスト　エラー数

エラー数							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態	全エラー
実験群	*M*	3.71	0.35	0.24	0.76	0.12	0.00	5.18
n=17	*SD*	2.42	0.59	0.42	0.64	0.32	0.00	4.40
比較群	*M*	4.13	0.13	0.13	0.56	0.06	0.00	5.00
n=16	*SD*	2.32	0.33	0.33	0.86	0.24	0.00	4.08
統制群	*M*	3.43	0.21	0.07	0.29	0.07	0.00	4.07
n=15	*SD*	3.09	0.56	0.26	0.59	0.26	0.00	4.75

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表4: 事前テスト　発生件数

発生件数							
		過去形	動詞	目的節	前置詞	能動態	受動態
実験群	*M*	8.12	8.12	1.29	7.53	3.76	0.06
n=17	*SD*	3.16	3.16	1.36	2.93	1.86	0.24
比較群	*M*	10.13	10.13	1.81	10.63	5.44	0.06
n=16	*SD*	3.64	3.64	1.59	3.46	2.45	0.24
統制群	*M*	8.71	8.71	1.93	8.86	4.43	0.64
n=15	*SD*	2.76	2.76	1.10	3.27	2.47	2.06

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表5: 事前テスト　エラー率

エラー率（エラー数／発生件数）							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態
実験群	*M*	0.46	0.05	0.04	0.23	0.01	NA
n=17	*SD*	0.23	0.07	0.07	0.25	0.03	NA
比較群	*M*	0.45	0.01	0.01	0.12	0.01	NA
n=16	*SD*	0.26	0.04	0.03	0.17	0.02	NA
統制群	*M*	0.41	0.02	0.01	NA	0.01	NA
n=15	*SD*	0.33	0.06	0.05	NA	0.03	NA

（注）*M*は平均を表し、*SD*は標準偏差を表す。
（注）NAは、発話において文法項目の義務的文脈の発生がなく、計算から除外したものを示す。

おける多変量分散分析の結果，有意差はみられなかった（資料6）。2つのテストにおけるエラー率の差（表5, 14）は，動詞の過去形の使用において，3群とも減少し（図1），有意傾向があり，また小程度の効果量があった（$F(1, 88) = 3.77$, $p < .05$ $\eta^2 = .04$）。しかし，群間の差は有意差が認められず，効果量も確認されなかった（$F(2, 88) = .10$, $p < .90$ $\eta^2 = .00$）。動詞欠落のエラーは，

■表6: フィードバック後テスト　エラー数

エラー数							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態	全エラー
実験群	*M*	2.35	0.12	0.06	0.76	0.06	0.12	3.47
n=17	*SD*	2.06	0.32	0.24	0.81	0.24	0.47	4.13
比較群	*M*	1.38	0.19	0.00	0.31	0.06	0.00	1.94
n=16	*SD*	1.22	0.39	0.00	0.46	0.24	0.00	2.31
統制群	*M*	3.57	0.14	0.14	0.79	0.00	0.14	4.79
n=15	*SD*	3.64	0.52	0.52	0.77	0.00	0.52	5.96

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表7: フィードバック後テスト　発生件数

発生件数							
		過去形	動詞	目的節	前置詞	能動態	受動態
実験群	*M*	9.76	9.82	2.18	9.71	5.18	0.24
n=17	*SD*	2.24	2.18	1.76	2.40	2.38	0.55
比較群	*M*	10.63	10.63	2.69	10.69	4.94	0.25
n=16	*SD*	3.37	3.37	1.36	3.72	2.49	0.56
統制群	*M*	10.07	10.07	2.07	10.79	5.86	0.29
n=15	*SD*	2.52	2.52	0.80	2.62	3.04	0.59

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表8: フィードバック後テスト　エラー率

エラー率（エラー数／発生件数）							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態
実験群	*M*	0.23	0.01	0.01	0.24	0.00	NA
n=17	*SD*	0.17	0.03	0.02	0.32	0.01	NA
比較群	*M*	0.12	0.02	0.00	0.05	0.00	NA
n=16	*SD*	0.11	0.04	0.00	0.09	0.02	NA
統制群	*M*	0.34	0.01	0.02	0.12	0.00	NA
n=15	*SD*	0.33	0.05	0.06	0.13	0.00	NA

（注）*M*は平均を表し、*SD*は標準偏差を表す。
（注）NAは、発話において文法項目の義務的文脈の発生がなく、計算から除外したものを示す。

実験群のみ減少し，比較群および統制群においては変化がみられなかった。また，群間の差は多変量分散分析により有意差が認められ，また中程度の効果量も確認された（F（2, 88）= 2.84, p < .06 η^2 = .06）。目的節欠落のエラーは，実験群のみ減少し，比較群と統制群では変化がみられなかった。しかし，多変量分散分析の結果，テスト間（F（1, 88）= 1.02, p < .31 η^2 = .01）およ

■表9: 直後事後テスト　エラー数

エラー数							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態	全エラー
実験群	*M*	3.24	0.12	0.00	0.82	0.06	0.00	4.24
n=17	*SD*	2.16	0.32	0.00	1.04	0.24	0.00	3.76
比較群	*M*	3.06	0.06	0.06	0.63	0.00	0.00	3.81
n=16	*SD*	2.25	0.24	0.24	0.78	0.00	0.00	3.51
統制群	*M*	3.71	0.14	0.07	0.79	0.00	0.14	4.86
n=15	*SD*	2.43	0.35	0.26	0.86	0.00	0.35	4.25

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表10: 直後事後テスト　発生件数

発生件数							
		過去形	動詞	目的節	前置詞	能動態	受動態
実験群	*M*	8.82	8.82	2.00	8.65	4.29	0.24
n=17	*SD*	3.09	3.09	1.64	2.27	1.71	0.42
比較群	*M*	9.69	9.69	2.00	10.00	5.25	0.19
n=16	*SD*	2.31	2.31	1.54	3.64	2.25	0.53
統制群	*M*	9.36	9.36	2.36	9.36	5.00	0.29
n=15	*SD*	2.44	2.44	1.63	3.04	2.45	0.45

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表11: 直後事後テスト　エラー率

エラー率（エラー数／発生件数）							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態
実験群	*M*	0.35	0.02	0.00	0.21	0.01	NA
n=17	*SD*	0.22	0.05	0.00	0.29	0.02	NA
比較群	*M*	0.33	0.01	0.00	0.12	0.00	NA
n=16	*SD*	0.28	0.03	0.02	0.18	0.00	NA
統制群	*M*	0.41	0.01	0.01	0.23	0.00	NA
n=15	*SD*	0.32	0.03	0.03	0.28	0.00	NA

（注）*M*は平均を表し、*SD*は標準偏差を表す。
（注）NAは、発話において文法項目の義務的文脈の発生がなく、計算から除外したものを示す。

び群間（F（2, 88）= .57, p < .56 η^2 = .01）の有意差は認められず，小程度の効果量が確認された。前置詞の欠落は，実験群で減少した。比較群および統制群については，比較群は遅延事後テスト，統制群は事前テストにおいて発生がない発話があったため，テスト間の比較は行わなかった。また，能動態であるべき箇所を誤って受動態にするエラーは，実験群および統制群においてわず

■表12: 遅延事後テスト　エラー数

エラー数							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態	全エラー
実験群	*M*	2.59	0.29	0.12	0.53	0.00	0.00	3.53
n=17	*SD*	2.22	0.46	0.32	0.61	0.00	0.00	3.61
比較群	*M*	3.38	0.06	0.13	0.44	0.19	0.00	4.19
n=16	*SD*	3.12	0.24	0.33	0.79	0.39	0.00	4.87
統制群	*M*	3.14	0.14	0.14	0.57	0.00	0.07	4.07
n=15	*SD*	2.26	0.35	0.35	0.73	0.00	0.26	3.95

（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表13: 遅延事後テスト　発生件数

発生件数							
		過去形	動詞	目的節	前置詞	能動態	受動態
実験群	*M*	8.35	8.47	1.94	9.41	5.47	0.35
n=17	*SD*	3.01	3.07	1.51	3.36	2.93	0.48
比較群	*M*	9.38	9.38	2.50	9.44	5.44	0.19
n=16	*SD*	3.24	3.26	1.58	3.55	3.92	0.39
統制群	*M*	9.29	9.29	1.79	7.57	3.43	0.29
n=15	*SD*	2.28	2.28	1.52	3.33	1.99	0.59

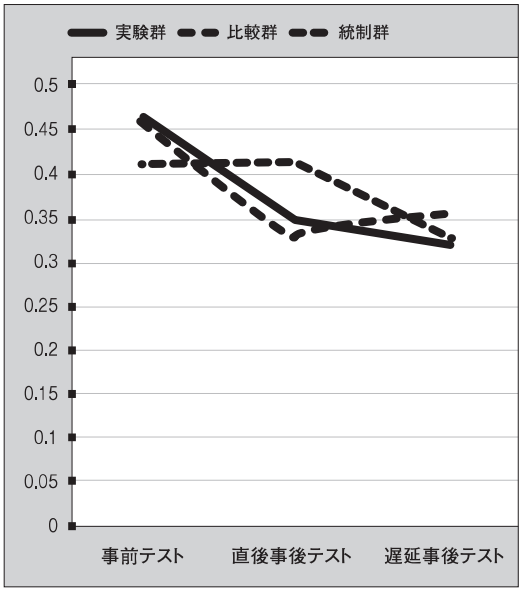
（注）*M*は平均を表し、*SD*は標準偏差を表す。

■表14: 遅延事後テスト　エラー率

エラー率（エラー数／発生件数）							
		過去形	動詞欠落	目的節欠落	前置詞欠落	能動態→受動態	受動態→能動態
実験群	*M*	0.32	0.04	0.01	0.08	0.00	NA
n=17	*SD*	0.25	0.06	0.03	0.10	0.00	NA
比較群	*M*	0.35	0.01	0.01	NA	0.02	NA
n=16	*SD*	0.29	0.03	0.03	NA	0.04	NA
統制群	*M*	0.33	0.02	0.01	0.16	0.00	NA
n=15	*SD*	0.24	0.04	0.03	0.21	0.00	NA

（注）*M*は平均を表し、*SD*は標準偏差を表す。
（注）NAは、発話において文法項目の義務的文脈の発生がなく、計算から除外したものを示す。

かな現象があったが，テスト間（$F(1, 88) = .18$, $p < .66$ $\eta^2 = .00$）および群間（$F(2, 88) = .88$, $p < .41$ $\eta^2 = .01$）において有意差はなかった。テスト間においては効果量がなかったが，群間では小程度の効果量が確認された。



■図1：3群の動詞過去形のエラー率

4 議論

4.1 議論と考察

実験の結果から，全体における比較，焦点化直接的修正フィードバックの対象であった動詞過去形の使用に関するエラー率の比較，また非焦点化直接的修正フィードバックの対象であった定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラー率の比較を行う。

4.1.1 直後事後テストにおける焦点化直接的修正フィードバックの影響

研究課題1は，動詞の過去形のエラーについて焦点化直接修正フィードバックを与えた場合，非焦点化直接修正フィードバックを与えた場合と比して，効果事前テストと直後事後テストの動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラー率が減少するのかであった。発生件数に対するエラーの割合は，6つの文法項目を総合的に見ると，実験群および比較群において減少したが，統制群では差が確認されなかった。多変量分散分析の結果によるとテスト間に有意差があり，群間には有意差が認められなかったため，直接的修正フィードバックに関わらず，文法的正確性に意識が向けられている可能性が示唆された。

しかし，事前テスト，フィードバック後のテスト（事前テストと同じタスク），直後事後テスト，遅延事後テストにおける文法項目別に3群のエラー率の変化を見ると，異なる傾向が確認された。動詞の過去形において，事前テストとフィードバック後のテストのエラー率の減少は群間で差がみられた。事前テストでは，実験群の過去形のエラー率（.46）は比較群（.45）と差がなかったが，事前テストの発話のスクリプトおよびフィードバックを読んだ後に同じタスクを行ったとき，実験群の過去形のエラー率（.23）は比較群のエラー率（.12）より高かった。実験群は動詞の過去形のエラーのみに修正フィードバックを与えられたにもかかわらず，比較群と比してエラー率は減少しなかった。また，動詞の過去形を使用する義務的文脈の件数は，実験群（事前テスト 平均8.12件，フィードバック後のテスト 平均9.76件）も比較群（事前テスト 平均10.13件，フィードバック後のテスト 平均10.63件）も減少していないため，過去形の使用の回避はみられなかった。更に，統制群の事前テストの過去形のエラー率（.41）は実験群や比較群より少なかったが，スクリプトを読んだ後に同じタスクを再度行ったとき，統制群は他の2群と比してエラー率が高かった（.34）。この結果から，文法的正確性について直接的フィードバックを行った場合，文法的正確性に注意が向き，動詞の過去形に関しては正確性が向上することが示唆された。しかし，その後に直後事後テストとして別のタスクを行った場合，群間で有意差はなかったため，新しいタスクの実施時の発話には修正フィードバックの影響は及ばないことがわかった。

目的節の欠落については，群間で有意差がみられた。エラー率の数値は小さくエラーの変化を観測するのが困難であるため，エラー数と義務的文脈の件数を検証したところ，実験群と比較群において，事前テストにおける目的節欠落のエラー

数は，統制群より多かった（実験群 平均.24回，比較群 平均.13回，統制群 平均.07回）が，フィードバック後に同じタスクを行ったとき，実験群と比較群の発話における目的節の義務的文脈の件数は増加しているにもかかわらず，エラー数は減少した（実験群 平均.06回，比較群 平均.00回）。その一方で，統制群のエラー数は増加した（平均.14回）。しかし直後遅延テストでは，実験群の発話におけるエラー数は減少したが（平均.00回），比較群の発話ではエラー数が増加し（平均.06回），統制群では事前テストと同様の結果となった（平均.07回）。直接的修正フィードバックを与えられた後に同じタスクを行ったときの発話において，文法的側面への注意が増加して目的節の欠落のエラーは減少するが，動詞により目的節の必要性が左右されるため，目的節の使用は，文法的側面だけでなく語彙的側面も関係しており，新しいタスクを行ったときには修正フィードバックの効果が表れなかったと考えられる。

前置詞の欠落のエラーは，統制群の事前テストの発話では義務的文脈がみられなかったため，実験群と比較群の発話のみ検証した。事前テストにおいて，実験群のエラー率は.23，比較群は.12であったが，フィードバック後のテストにおいて，実験群のエラー率は変化がみられなかった（.24）のに対し，比較群は大きく減少した（.5）。しかし，直後事後テストで新しいタスクを行ったとき，エラー率は事前テストと変わらない結果となった（実験群 .21，比較群 .12）。前置詞の正確な使用には，後続する目的語に関する語彙的知識が必要である。事前テストで誤って使用した前置詞について直接的修正フィードバックを与えられた比較群は，フィードバック後の同じタスクにおいて正しく使用することが可能となりエラー率が減少したが，その後に新しいタスクを行ったときは事前テストとは別の目的語を使用し，それらの目的語における前置詞に関するインプットを受けていなかったため，前置詞の正確な使用において事前テストと同様の結果になったと考えられる。

これらの結果から，直接的修正フィードバックは，焦点化されている場合および焦点化されていない場合において，フィードバック後のテストおよび直後事後テストの発話の正確性が向上する場合があり，文法的正確性に注意が向けられていることが示唆されたが，語彙的知識を必要とする文法項目においては，修正フィードバックの効果はみられないことが明らかになった。

4.1.2 遅延事後テストにおける焦点化直接的修正フィードバックの影響

研究課題2は，動詞の過去形のエラーについて焦点化直接修正フィードバックを与えた場合，非焦点化直接修正フィードバックを与えた場合と比して，効果事前テストと遅延事後テストの動詞の過去形，定動詞の欠落，目的語の欠落，前置詞の欠落，能動態および受動態のエラー率が減少するのかであった。

多変量分散分析の結果，動詞過去形のエラーに関して，テスト間の差に有意傾向がみられた。実験群と比較群の発話では，フィードバック後のテストにおいて事前テストからエラー率が減少したのに対し，統制群では大きな減少がみられなかった（実験群 .46→.23，比較群 .45→.12，統制群 .41→.34）。しかし，3，4週間後に行った遅延事後テストの発話では，3群のエラー率に差がみられなかった（実験群 .32，比較群 .35，統制群 .33）。この結果から，直接的修正フィードバックの有無に関わらず，漫画描写タスクを繰り返し行い，またその内の1つの発話を書き起こしたスクリプトを読むことにより，動詞の過去形のエラーは減少することが示唆された。

また，目的節の欠落に関して，群間に差がみられた。実験群の発話において，事前テストおよび遅延事後テストのエラー率は減少した（.04→.01）が，比較群および統制群の発話ではテスト間に差がみられなかった（比較群 .01→.01，統制群 .02→.02）。エラー数をみると，事前テストにおいて実験群は4名の発話において目的節欠落のエラーがみられたが，遅延事後テストではエラーがあったのは2名の発話となっていた。その一方で，比較群では事前テストで2名の発話においてエラーがあり，遅延事後テストにおいても2名の発話でエラーがみられた。また，統制群では，事前テストでは1名の発話でエラーがあったが，遅延事後テストでは2名の発話でエラーが発生した。能動態を使用すべき箇所において受動態を使用するエラーは，事前テストにおいて各群ともエ

ラーが発生していたが(実験群 平均 .12回, エラー率 .01, 統制群 平均 .07回, エラー率 .01), 遅延事後テストにおいては, 実験群と統制群ではほとんどエラーがみられなくなった(実験群 平均.00回, エラー率 .00, 統制群 平均 .00回, エラー率 .00)。その一方で, 比較群はエラーが増加した(事前テスト 平均 .06回, エラー率 .01, 遅延事後テスト 平均 .19回, エラー率 .02)。比較群は, 修正フィードバックの対象文法項目が他の2群と比して多かったために, 態の正しい使用に注意を向けることが困難になった可能性があるが, 3群の発話において, 態のエラーの発生件数が少ないため, 更なる検証が必要である。

これらの結果から, 直接的修正フィードバックは, 焦点化されている場合および焦点化されていない場合において, フィードバック後のテストおよび遅延事後テストの発話の正確性は, 文法項目により若干の差がみられるものの, 大きな差はみられないことが示唆された。動詞の過去形のエラーについては, 直接的修正フィードバックは, 遅延事後テストには影響が及ばず, 漫画描写タスクを複数回行うことや, 発話を書き起こしたスクリプトを読むことが文法的正確性の長期的な向上につながると考えられる。しかし, 今回の結果は, 漫画描写タスクの繰り返しの影響によるものか, 発話のスクリプトを読んだことによるものか検証していないため, これらの2つの要因について検証が必要である。

4.1.3 直接的修正フィードバックの効果

本研究は, 焦点化および非焦点化直接的修正フィードバックの影響を検証したが, 2種類の修正フィードバックの効果には大きな差がみられなかった。直接的修正フィードバックを与えた文法項目のエラー率が減少するとは限らず, 修正フィードバックを与えなかった文法項目のエラー率も減少した。また, 直接的修正フィードバックを与えた実験群と比較群は, 直接的修正フィードバックを与えなかった統制群と比して, 全体的なエラー率が減少した。これらの結果から, 修正フィードバックを与えることは文法的正確性に注意を向けることを促し, 文法的エラーが減少する効果があると考えられる。しかし, 多変量分散分析の結果, 有意差がみられない項目が多かったため, 解釈には注意が必要である。

また, 目的節や前置詞は, 文法的知識だけでなく語彙的知識も必要とする。そのため, 文法的正確性に注意を向けるための修正フィードバックのみでエラーを防ぐことが難しく, 語彙に関する指導も必要である。

Truscott (2004) は, 第二言語のライティング指導において, 修正フィードバックを行うことにより, フィードバック対象である文法項目の使用を避ける可能性があると指摘しているが, スピーキング指導においてはその限りではないと言える。本研究において, 実験参加者が直接的修正フィードバックの対象文法項目の使用を避ける傾向はみられなかった。修正フィードバックを与えられた実験群と比較群の発話における全てのフィードバック対象文法項目について, 文法項目を使用しなければならない義務的文脈の件数は減少していないことから, 修正フィードバックを受けても当該文法項目の回避は起こらないことが明らかになった。これは, 文法的エラーが起きたときに修正フィードバックを与えることにより対象文法項目の使用を避けて練習の機会を失う可能性について懸念する必要がないと言えるであろう。

4.1.4 スピーキングとライティングにおける焦点化直接的修正フィードバックの効果

第二言語ライティングにおける焦点化直接的修正フィードバックに関する研究では, 焦点化および非焦点化直接的修正フィードバックを受けた場合, 直接的修正フィードバックを受けなかった場合と比して, 新しいタスクを行う直後事後テストにおいて焦点化修正フィードバックの対象文法項目である冠詞の正確性が向上した。また, 遅延事後テストでは, 焦点化直接的修正フィードバックを受けた群は, 非焦点化直接的修正フィードバックを受けた群と比して冠詞の正確性が向上した(Ellis, et al., 2008)。本研究における第二言語スピーキングに関する調査では, ライティングの先行研究と同様に, 直後事後テストにおいて, 修正フィードバックを受けた実験群と比較群の発話では, 焦点化修正フィードバックの対象文法項目である動詞の過去形の正確性が向上したが, 遅延事後テストにおいても, 修正フィードバックを受けた2群は動詞の過去形の正確性の向上が

あり，実験群と比較群の間に差はみられなかった。この結果より，少なくともスピーキング指導において，焦点化直接的修正フィードバックも非焦点化直接的修正フィードバックも，同等の効果があるといえるであろう。

4.2 今後の課題

本研究は，スピーキング指導における直接的修正フィードバックを扱った。口頭ではなく筆記によるフィードバックを行ったが，その理由は，2.3の本研究で実施するフィードバックの項で述べたように，実験目的に合ったフィードバックを正確に実施するために，筆記によるフィードバックを行ったが，その結果，発話を書き起こしたものを実験参加者が読むという要因が発生し，それが結果に影響を及ぼした可能性がある。

また，本研究はスピーキング練習における直接的修正フィードバックであるが，口頭で直接的修正フィードバックを与えずに，書き起こしたものを見せて直接的修正フィードバックを与えた。そのため，通常の教室のスピーキング指導において多くみられるフィードバックとは異なるものであり，直接応用することは難しいであろう。しかし，筆記による直接的修正フィードバックは，大きなクラスでは現実的なフィードバック方法であり，本研究の結果は，学校教育において多くみられる，30名や40名といった大教室で行われる英語スピーキングの授業では，指導への示唆となるであろう。

焦点化された直接的修正フィードバックを与える文法項目は，本研究では動詞の過去形を対象とした。しかし，実験参加者の発話において動詞の過去形のエラーは最も多く発生しており，その結果，事前テスト後に行ったフィードバックの回数は，実験群は平均3.71回（SD=2.42），比較群は平均5.00回（SD=4.08）であり，実験群と比較群に対して与えたフィードバックの回数は十分な差がなかった可能性があり，そのため焦点化と非焦点化直接的修正フィードバックの影響が十分にみられなかったと考えられる。フィードバックを実施する文法項目を変えて検証する必要がある。

実験参加者の熟達度が結果に影響を及ぼした可能性も考慮すべきであろう。本研究では，参加者は無作為に3群に分けられたため，各群のTOEFL ITP® の点数には差が発生した（実験群 M = 451.611, SD = 167.917, 比較群 M = 378.944, SD = 204.911, 統制群 M = 470.563, SD = 128.163）。また，全参加者の熟達度においても差がみられた（Max = 600, Min = 393）。そのため，修正フィードバックを受けたとき，フィードバックの内容を正しく理解し，理解したことをその後の発話に応用する能力において，参加者間で差があった可能性がある。参加者の熟達度が結果に影響を及ぼした可能性があるため，熟達度を統制して追調査を行う必要がある。

また，動詞の過去形以外の文法項目において，エラーの発生件数が少なかったことも結果に影響を及ぼしていると考えられる（実験群 平均3.71回，比較群 平均5.00回）。比較群には非焦点化直接修正フィードバックを与えることにより，修正フィードバックの対象を増やし，実験参加者の注意を分散することを意図して実験をデザインしたが，実験群と比較群のフィードバック件数に十分な差がなかった可能性がある。更に，エラーの発生件数が少なかったため，十分な分析が行われなかったことも結果に影響していると考えられる。

実行可能性の問題から，本研究では実験参加者の人数が限られていたが，今後の課題としてサンプルサイズが少なかったことが挙げられる。サンプルサイズを増やし，より正確な調査が必要であろう。

謝辞

本研究の機会を与えてくださいました公益財団法人 日本英語検定協会の皆様と，研究助成選考委員の先生方，特に，助言担当の池田央先生に心より感謝申し上げます。また，名古屋大学の木下徹教授には研究に関する示唆に富んだご助言をいただきました。この場をお借りして感謝申し上げます。最後に，調査実施に際しては，調査に協力して頂いた48名の大学学部生の皆様に厚く御礼申し上げます。

参考文献（＊は引用文献）

AC Del Re (2013). compute.es: Compute Effect Sizes. R package version 0.2-2. URL http://cran.r-project.org/web/packages/compute.es

＊ Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267-296. doi:10.1016/S1060-3743(03)00038-9

＊ Date, M., & Takatsuka, S. (2012). Effects of Task Repetition on Oral Performance of the Same Task and a New Task of the Same Type. *ARELE, 23*, 345-360. DOI: 10.20581/arele.23.0 345

＊ Ellis, N. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. Studies in *Second Language Acquisition, 27*, 305-352. 10+10170S027226310505014X

＊ Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*, 358-371. 10.1016/j.system.2008.02.001

Hadley Wickham, Romain Francois, Lionel Henry and Kirill Muller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr

Hays, M. J., Kornell, N. & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 290-296.

Heaton, J. (1975). Beginning composition through pictures. Harlow: Longman.

＊ Kartchava, E. & Ammar, A. (2014). The noticeability and effectiveness of corrective feedback in relation to target type. *Language Teaching Research, 18*(4), 428-452. 10.1177/1362168813519373

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics, 27*(3), 405-430.

Panova, I. & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ELS classroom. *TESOL Quarterly, 36*(4), 573-595.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version= 1.8.3.

＊ Saeb, F. (2014). The effects of focused and unfocused written corrective feedback on the grammatical accuracy of beginner EFL learners. *International Journal of Applied Linguistics & English Literature, 3*(2), 22-26. 10.7575/aiac.ijalel.v.3n.2p.22

＊ Sheen, Y. (2007). The effect of focused written corrective feedback and languagage aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*, 255-283.

＊ Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*. 556-569. DOI: 10.1016/j.system.2009.09.002

＊ Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of Written Feedback and Revision on Learners' Accuracy in Using Two English Grammatical Structures. *Language Learning, 64*(1). 103-131. DOI: 10.1111/lang.12029

＊ 白畑知彦 (2015).『英語指導における効果的な誤り訂正――第二言語取得研究の見地から』東京：大修館書店

Slawomir Jarek (2012). mvnormtest: Normality test for multivariate variables. R package version 0.1-9. https://CRAN.R-project.org/package=mvnormtest

＊ Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369.

＊ Truscott, J. (1999). The case for " The case against grammar correction in L2 writing classes " : A response to Ferris. *Journal of Second Language Writing, 8*(2), 111-122.

＊ Truscott, J. (2004). Evidence and conjecture on the effects of correction : A response to Chandler. *Journal of Second Language Writing, 13*, 337-343. 10.1016/j.jslw.2004.05.002

＊ Van Beuningen, C., De Jong, N., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning, 62*(1). 1-41. DOI: 10.1111/j.1467-9922.2011.00674.x

Wickham, H. [aut, cre] & RStudio [cph], (2016). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions* (R package version 0.6.0). Retrieved from https://CRAN.R-project.org/package=tidyr

資料1：事前テスト・直後事後テスト

	自由度	Pillai	approx F	要因の自由度	残差の自由度	Pr(>F)
テスト	1	0.112858	2.70332	4	85	0.03572*
群	2	0.062141	0.68944	8	172	0.70050
テスト・群	2	0.054591	0.60332	8	172	0.77427
残差	88					

(注)0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料2：事前テスト・直後事後テスト，過去形

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.1563	0.156344	1.9894	0.1619	0.02187
群	2	0.0055	0.002753	0.0350	0.9656	0.00077
テスト・群	2	0.0704	0.035191	0.4478	0.6405	0.00985
残差	88	6.9159	0.078589			
計	93	7.1481				

(注)0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料3：事前テスト・直後事後テスト，動詞欠落

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.005890	0.0058900	2.3535	0.1286	0.02504
群	2	0.006822	0.0034108	1.3629	0.2613	0.02900
テスト・群	2	0.002292	0.0011461	0.4579	0.6341	0.00974
残差	88	0.220232	0.0025026			
計	93	0.235236				

(注)0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料4：事前テスト・直後事後テスト，目的節欠落

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.006390	0.0063896	3.6884	0.05803	0.03871
群	2	0.001623	0.0008116	0.4685	0.62751	0.00983
テスト・群	2	0.004609	0.0023045	1.3303	0.26967	0.02792
残差	88	0.152448	0.0017324			
計		0.16507				

(注)0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料5：事前テスト・直後事後テスト，能動態→受動態

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.000893	0.00089250	1.8167	0.1812	0.02002
群	2	0.000472	0.00023621	0.4808	0.6199	0.01058
テスト・群	2	0.000014	0.00000675	0.0137	0.9863	0.00031
残差	88	0.043233	0.00049128			
計	93	0.044612				

(注)0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料6： 事前テスト・遅延事後テスト

	自由度	Pillai	approx F	要因の自由度	残差の自由度	Pr(>F)
テスト	1	0.055511	1.24893	4	85	0.2966
群	2	0.099268	1.12287	8	172	0.3502
テスト・群	2	0.057117	0.63206	8	172	0.7501
残差	88					

(注) 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料7： 事前テスト・遅延事後テスト, 過去形

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.2878	0.287763	3.7773	0.05515	0.04099
群	2	0.0161	0.008033	0.1054	0.90003	0.00229
テスト・群	2	0.0131	0.006569	0.0862	0.91747	0.00187
残差	88	6.7040	0.076182			
計	93					

(注) 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料8： 事前テスト・遅延事後テスト, 動詞欠落

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.000815	0.0008152	0.2681	0.60592	0.00285
群	2	0.017275	0.0086374	2.8403	0.06377	0.06046
テスト・群	2	0.000028	0.0000142	0.0047	0,99534	0.00010
残差	88	0.267607	0.0030410			
計	93	0.285725				

(注) 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料9： 事前テスト・遅延事後テスト, 目的節欠落

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.002050	0.0020502	1.0205	0.3152	0.01111
群	2	0.002302	0.0011509	0.5729	0.5660	0.01248
テスト・群	2	0.003370	0.0016848	0.8386	0.4357	0.01827
残差	88	0.176783	0.0020089			
計	93	0.184505				

(注) 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

資料10： 事前テスト・遅延事後テスト, 能動態→受動態

	自由度	平方和	平均平方	統計量	Pr(>F)	η^2
テスト	1	0.000120	0.00012024	0.1891	0.6648	0.00202
群	2	0.001131	0.00056574	0.8896	0.4145	0.01905
テスト・群	2	0.002144	0.00107225	1.6861	0.1912	0.03612
残差	88	0.055963	0.00063594			
計	93					

(注) 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

第30回 研究助成
A 研究部門 ● 報告Ⅴ ● 英語能力テストに関する研究

パラフレージング・レベルによる錯乱肢の作成

― 誤答分析に基づくテキスト理解度の診断に向けて ―

研究者:茨城県／筑波大学大学院　在籍　政所 里佳
《研究助言者:村木 英治》

概要　本研究では，パラフレーズ質問の選択肢における言い換えの程度について，原文の語句を含む割合であるパラフレージング・レベルの観点から，現状の調査および実証研究を行った。調査1では現行の読解テストのパラフレーズ質問の選択肢を対象に，記述と英文中の語が重複している程度について比較した結果，パラフレージング・レベルが低い選択肢の割合はテスト間で異なることがわかった。調査2では，パラフレージング・レベルに基づいて選択肢を作成した読解問題を実施した結果，正答・誤答選択肢の組み合わせによる選択率と魅力度への影響は，テスト得点群によって異なることが示された。テスト得点がより低い受験者は，正答選択肢のパラフレージング・レベルが低い場合に，よりレベルが高い錯乱肢を多く選択し，魅力的だと判断する傾向が見られた。結果より，パラフレーズ質問の選択肢作成には，文章中の記述を含む割合や選択肢間の関係性を併せて検討することが重要であると示唆された。

1 はじめに

多肢選択式問題は，今日までの教育的測定において，最も有効かつ長く使用されている形式の一つである。その理由として，採点の容易さと客観性が確保でき，高い信頼性が期待できるなどの長所があることが挙げられる。このような利点から英語教育分野でも，英検をはじめとする多くの大規模テストにおいて多肢選択式問題が出題されてきた。しかし，多肢選択式問題には多くの問題点も指摘されている。中でも主要な問題の一つに，当て推量による正解の可能性が否定できないことが挙げられる。つまり，4つの選択肢から1つを選ぶこの形式には，受験者が問題で求められる理解を達成できているか否かに関係なく，4分の1の確率で正解してしまう恐れがあるということである。

このような当て推量による正解を防ぐ方策として，これまでの研究では，良い誤答の選択肢，すなわち，能力のある受験者には魅力的には見えないが，能力のない受験者にとっては思わず選んでしまう（根岸，2017）ような錯乱肢をどのように作成すればよいのかという問いが検証されてきた。しかしながら，このような効果的な錯乱肢の性質について検証した研究からは，そのもっともらしさを構成する要素について，十分に一貫した結果が得られているとは言い難い。その理由として，各研究において対象とした問題の質問タイプやテキストなどの要因が異なっていることが挙げられる。また，複数の選択肢から最終的に1つの選択肢に絞る受験者の解答プロセスを考慮すると，正答選択肢と錯乱肢それぞれの性質のみならず，両者の関係性を含めた検証によって，受験者のパフォーマンスをより詳細に診断できると考えられる。しかし，そのような試みを行っている研究はこれまで多くは見られなかった。

そこで，本研究では，多肢選択式問題読解テス

トにおいて大きな割合を占める質問タイプであるパラフレーズ質問に焦点を当て，記述中にテキスト原文にある語を含む割合であるパラフレージング・レベルに基づく錯乱肢を作成し，その有用性について，正答選択肢との関係や選択肢の魅力度の観点から検証する。テキスト理解が十分に達成されている受験者にとっては，テキスト原文の語を含む割合に関係なく，自信をもって正答選択肢を選ぶことができると考えられる。一方，十分に理解しきれていない受験者の場合は，テキスト原文の語を含む割合の高低が，選択肢の判断を迷わせる可能性が考えられる。

本研究は2つの調査から構成される。調査1では，現行の読解テストのパラフレーズ質問を対象に，パラフレージング・レベルの観点から，正答・誤答の選択肢が含む語と英文中の語はどの程度重複しているのかについて調査を行った。続く調査2においては，パラフレージング・レベルに基づいて正答選択肢・錯乱肢を作成したテストを用いて実験を実施した。結果からその有用性を検討し，弁別力のあるパラフレーズ質問の錯乱肢の作成方法に示唆を与えることを試みた。

2 先行研究

2.1 読解テストにおけるパラフレーズ質問

リーディングテストは，出題形式や質問項目，正答・誤答の選択肢の要因によって，測定される能力が変化する。

例えば，清水（2005）は英検，TOEFL，大学入試センター試験問題の問題を分析し，リーディングテストにおいて用いられる質問タイプは大きく次の6つに分けられることを示した（表1）。

このように，質問タイプによって解答者に求められるテキストの理解は異なっている。つまり，テスト作成者は各質問タイプが読み手のどのような能力を測っているのかを把握した上で，どのタイプを用いるかを決める必要がある。

一般に，読解において達成される「理解の深さ」については3つのレベルがあるとされる（van Dijk & Kintsch, 1983）。テキストを読んだ読み手の心内に記憶される痕跡は「心的表象」と呼ばれ，そのうち（a）表層的記憶（テキスト内の語や文をそのまま記憶している段階），（b）命題的テキストベース（テキストに含まれる命題の理解），（c）状況モデル（読み手の知識や推論を含めて構築された表象）の順に，テキストの理解は達成される。

したがって，例えば，テキストの中に出現する単語や文がそのまま正答選択肢に含まれる質問を出題した場合，表層的記憶とほぼ同じレベルの理解を測っているといえる。一方，選択肢に含まれる語がテキストに含まれる単語や文と全く同じでない場合，すなわち，単にテキスト内の単語や文の異同を問うものではないパラフレーズ質問や指示質問においては，読み手がテキストの中に含まれる命題をどれだけ理解しているかという命題的テキストベースの理解を測っているといえる（卯城，2012）。

清水（2005）において明らかになったのは，大規模読解テストにおけるパラフレーズ質問の多さである。特に，英検の長文読解問題においては，およそ9割の設問がこの質問タイプに該当することが示されている。このようにパラフレーズ質問が多くの読解テストにおいて採用されている理由の一つとして，質問作成の容易さ，解答の作成のしやすさから，作成者にとって作りやすいこ

■表1：読解テストにおける質問タイプ（清水，2005）

- パラフレーズ質問（文章中の局所的なある一部分を言い換えると，質問とそれに対する正解が得られる質問。明示情報に関する質問）
- 推論質問（文章に基づいて適切に推論されることについて問う質問）
- テーマ質問（パラグラフまたは文章全体の主題について問う質問）
- 提示質問（代名詞または指示表現の先行詞の理解について問う質問）
- 語彙問題（語彙の意味について問う質問）
- 文章構造質問（比較・対照や時間順など文章構造について問う質問やある内容が文章中のどの部分で述べられていたかを問う質問）

とが挙げられる。また，清水は，質問文や選択肢を言い換えることで，文章中のある部分とおおよそ一致するパラフレーズ質問は，受験者自身でパラフレーズする必要はあるが文章中に明示されている情報について解答することを求めるため，下位レベル処理の質問に分類されるとした。

このように，項目作成が容易で困難度がそれほど高くないことから，質問タイプの中でも出題や解答の機会が最も多いと考えられるパラフレーズ質問であるが，未だに明らかにされていない点がいくつかある。

第一に，定義としては同じ「パラフレーズ質問」という分類に該当する項目であっても，選択肢の言い換えの程度は統制されておらず，項目間で異なることが多い。例えば，清水(2005)の行っていた設問分類では，パラフレーズ質問の中に，テキスト内の単語や文をそのまま字義通りに解答すれば正答となる字義的な質問も含まれていた。つまり，パラフレーズ質問における選択肢が，テキスト内の単語や文を言い換えている程度，含んでいる割合について厳密な場合分けを行った試みは未だに十分であるとは言い難い。前述したように，質問項目の選択肢がテキスト内の単語や文をどの程度含んでいるのかによって，読み手のテキストの理解のうち，どの段階を測っているのかが変わる可能性がある。構成概念の検証のためには，選択肢がテキスト内の単語や文を含む程度，言い換えている程度を考慮することが求められると考えられる。

第二に，項目の選択肢どうしの関係性である。前述したように，パラフレーズ質問ではテキストの一部分を言い換えることにより，質問と正答選択肢が作成される。しかし，正答選択肢以外の錯乱肢における，テキストに含まれる単語や文等の表現を含んでいる程度については検討する必要はないのだろうか。多肢選択式テストに解答する受験者の立場に立ってみると，複数の選択肢から1つの正答を選ぶ際，最初から1つの正答に絞って選ぶ場合を除けば，他の選択肢の内容も考慮して吟味すると考えられる。すなわち，正答選択肢と錯乱肢との関係性(e.g., 正答選択肢・錯乱肢においてテキスト内の語句を含んでいる程度)を考慮することは重要であると考えられる。

2.2 多肢選択式問題における錯乱肢の機能と性質

多肢選択式テストにおいて，通常，受験者は複数与えられた選択肢の中から1つの正答を選ぶ。この受験者に与える正答の選択肢や錯乱肢の作り方によって，問題の弁別力が変化する可能性があることから，選択肢の作成方法は問題作成者にとって重要な課題である。その中でも，錯乱肢は，理解が十分な受験者とそうでない受験者を弁別し，項目の有効性を高める上で重要な役割を担っている。

それでは，良い選択肢とはどのようなものなのか。Haladyna and Downing(1989)とHaladyna, Downing, and Rodriguez(2002)は，選択肢全般に関する多肢選択式問題作成のガイドラインを提示した。選択肢どうしの長さを均一に揃えることや，選択肢は類似したものとすることなど複数の選択肢を設ける上での留意事項に加え，錯乱肢については「もっともらしい錯乱肢であること」を提案している。また，多肢選択式テストの難易度を左右する要因を調査したDrum, Calfee, and Cook(1981)において，「錯乱肢のもっともらしさ」が難易度に有意に影響したことが示されている。加えて，Drum et al.(1981)は正答選択肢についても，低頻度語や新出内容語を含むか否かによって難易度が変化したことを指摘している。

このように，もっともらしい錯乱肢の重要性については明らかにされていたが，「もっともらしさ」に影響するのはどのような要因であるかについて，これまでその構成要素を調査した研究は多くなかった。

その中の一つに，Ushiro et al.(2007)がある。彼らは，読解テストにおける錯乱肢のもっともらしさの指標を調査するために，Drum et al.(1981)の指摘したテストの難易度を決定する要因のうち，(a)錯乱肢に含まれる語数，(b)錯乱肢に含まれる内容語の割合，(c)錯乱肢に含まれるテキスト内で用いられていない内容語の割合，(d)錯乱肢に含まれる否定語の総数，(e)錯乱肢に含まれる前置された語の総数，(f)錯乱肢に含まれる照応語の総数，の6つの要因について検証した。日本人大学生は，異なる質問タイプで構成されるTOEFLの多肢選択式問題に解答した。結果とし

て，質問タイプにより，これらの指標がもっともらしさの指標として機能するかどうかは異なっていた。特に，パラフレーズ質問においては，錯乱肢とテキストに含まれる語の重複が，もっともらしさに影響していたことが示されている。全体としては錯乱肢の表面的な性質は錯乱肢のもっともらしさとは関わりが弱いことが示されたものの，質問タイプによって選択肢のもっともらしさに影響する要因が異なる可能性が示唆されている。

もう一つの研究，Ushiro et al.（2008）では，大学生を対象とした実験で読解テストの質問タイプが自由記述式と多肢選択式の難易度に与える影響について検証し，多肢選択式では自由記述式とは異なり，項目難易度が質問タイプによって変化しないこと，そして選択肢の重要性が大きいことを示している。特に，難易度が高い項目ではどのような選択肢があったのか調査を行ったところ，いくつかの項目においてテキスト原文にある語と正答選択肢の語との重複の割合が非常に低い場合に正答率が下がっていたことが明らかになった。一方，錯乱肢の語とテキスト中の語の重複の割合については，項目によって一貫した結果は得られなかった。このことから，多肢選択式テストへの解答時には，受験者が選択肢から得られる情報に大きく頼っており，特にテキストと選択肢の語が重複しているものを選ぶ傾向があることが示唆されている。

また，髙木（2011）は，「記憶への保持されやすさ」という指標が錯乱肢の有用性を決定づけるかどうかについて調査を行った。具体的には，テキストの中における他の情報との因果的連続性の強弱を操作することで，記憶に保持されやすい錯乱肢と保持されにくい錯乱肢を作成し，日本人大学生を対象に，その効果を比較検証している。結果として，因果的連続性が強い錯乱肢は，テキストにない情報を記述した錯乱肢よりも，テキスト理解度の低い受験者を多く引き付けることを示した。

上記の研究結果を踏まえると，多肢選択式読解テストにおける選択肢のもっともらしさは，質問タイプや受験者のテキストの理解度により異なることが予想される。特に，前述した通り読解テストで最も多く用いられているパラフレーズ質問において，選択肢，とりわけ錯乱肢がテキストの情報を含む程度が受験者の解答パフォーマンスにどのような影響を与えるのかについては，検証の余地があると考えられる。したがって，本研究ではこの点について，次節で述べる手法を用いて調査を行うこととした。

2.3 パラフレージング・レベルを用いた多肢選択式問題

本節では，本研究が取り上げる2つの手法について述べる。まず，本研究で主要な選択肢の分類・作成基準として扱うのは，Keck（2006）のParaphrase Typesに基づいて考案した「パラフレージング・レベル」である。

Keck（2006）では，アメリカの大学生を対象に，L1とL2で書かれた1000語程度の文章をそれぞれ1パラグラフ程度におさめるように求める要約タスクを実施し，書き手がテキスト内で使用されている言葉を借用するストラテジーである「パラフレーズ」について調査を行っている。Keckは，書き手の行ったパラフレーズについて，元のテキストに含まれている情報がどの程度要約において複製されているか・言い換えられているかにより，次の4つの種類に分けた：(a) Near Copy, (b) Minimal Revision, (c) Moderate Revision, (d) Substantial Revision。さらに，Keckは，パラフレーズの割合を判断する基準として，「unique links（UL），general links（GL）」という枠組みを用いている。unique linksは，本文において局所的に表れる内容語（e.g.，名詞，形容詞，動詞），general linksは本文において広範囲に複数回表れる内容語であると定義しており，このULの語数の合計数が総語数に占める割合に準じて各パラフレーズが分類される（表2）。具体的には，合計数が50%以上であればNear Copy (NC), 20〜49%以上であればMinimal Revision (MIR), 1〜19%であればModerate Revision (MOR), 全く含まれない（0%）場合にはSubstantial Revision（SR）に分類されている。

本研究では，このKeck（2006）のParaphrase Typesの算出方法に関し，ULとGLの合計が選択肢の総語数に占める割合を求めるように一部修正したパラフレージング・レベルを基準として，多肢選択式読解テストにおけるパラフレーズ質

問の選択肢の分類を行うこととする。その理由は主に2点挙げられる。1点目に，本研究では表層的な語の性質に絞ってテキスト原文と選択肢の語の重複の程度の変化による影響を調査するため，重複する語の数が総語数に占める割合という，純粋かつ客観的な値を算出するKeckの枠組みが適していると考えるからである。2点目は，カウントする語の対象を内容語に絞っていることから，本研究でマテリアルとする読解テキストの機能語による影響を抑えられると考えたためである。

■表2: Paraphrase Typesの分類基準 (Keck, 2006)

	ULの語数の合計が総語数に占める割合
Near Copy (NC)	50%以上
Minimal Revision (MIR)	20 ～ 49%
Moderate Revision (MOR)	1 ～ 19%
Substantial Revision (SR)	0%

もう一つの手法として，本研究では，上記のパラフレージング・レベルを用いて作成された選択肢それぞれについて，受験者に対する魅力度を算出する。

多肢選択式の抱える課題は主に2つあるとされる。1つ目は，当て推量による正答の可能性があることである（Kobayashi, 2002）。つまり，内容を理解し，選択肢を吟味した上で正答を選んだ場合も，ランダムに選び4分の1の確率で当たった場合も，同じ「正解」となるため，作成者側からは適切な解答プロセスを経た解答とそうでない解答を区別することができない。2つ目は，解答者が正答を特定することはできなくても，いくつかの選択肢は不正解であることが分かっている場合に，その部分的知識を結果に反映させることができない点である。

このような多肢選択式テスト受験者の解答プロセスや部分的知識を反映させるために，様々な修正方法が考案されてきた。例えば，Shizuka (1999a, 1999b) では，受験者に解答自信度を3段階で評価させた値を，解答の正解・不正解と合わせて0～5の6段階で組み込む方法を採用すると，テストの信頼性や妥当性が向上したことを示している。この他，自信度に加え，解答までの反応時間や正答率を得点に反映させる方法もテストの信頼性向上を促したことが示されている（e.g., Shizuka, 2003）。

このような解答自信度の枠組みを応用し，Iimura (2014) では，正答選択肢と錯乱肢の，受験者に対するもっともらしさである「魅力度」を数値化した研究を行っている。Iimuraは熟達度の異なる日本人大学生を対象に，多肢選択式リスニングテストにおける選択肢への解答自信度を調査した結果から，受験者に選ばれる割合の低かった錯乱肢であっても，受験者は魅力的に感じている可能性があることを明らかにした。その上で，選択率などの解答頻度とは別の指標を用いて，各錯乱肢の魅力度を詳細に調べることが重要であるとしている。また，熟達度による違いとして，熟達度の低い学習者は錯乱肢に対してより高い魅力度を感じ，熟達度の高い学習者は正答に対する自信度がより高い傾向があることを明らかにしている。

以上のように，選択肢が受験者に与える影響を検証する上で，選択率といった解答頻度とは別に，受験者が各解答にどの程度自信をもったか，各選択肢をどの程度魅力的だと感じたかを調べるために，解答プロセスを表出化させることは重要であると考えられる。したがって，本研究では，パラフレージング・レベルを用いて作成された正答選択肢と錯乱肢それぞれの選択率に加え，魅力度を算出することにより，各選択肢が受験者にどのように作用するのかを検証する。

3 本研究の目的

本研究では，読解テストで最も多く用いられているパラフレーズ質問において，選択肢がテキストで使用される語を含んでいる程度が，テキスト理解度の異なる受験者の解答パフォーマンスに与える影響を調べることを目的とする。特に，記述の中にテキスト内の語を含んでいる割合であるパラフレージング・レベルを操作して作成された選択肢の有用性を検証するため，本研究では2つの調査を行った。まず調査1では，3つの現行の読解テストを対象に，各項目の選択肢のパラフレージング・レベルにはどのような傾向があるのかを調査した。続く調査2では，調査1から明らかになった傾向に基づいて作成された，異なるパラフレージング・レベルの正答選択肢・錯乱肢が受験者に選ばれた割合や魅力度を分析し，その有効性を検証した。

4 調査

4.1 目的

調査1の目的は，現在運用されている読解テストのパラフレーズ質問を対象とし，各項目の選択肢のパラフレージング・レベルの傾向を明らかにすることである。また，調査結果をテスト間で比較することにより，テストごとの選択肢の特徴およびその理由についても考察する。

4.2 マテリアル

現在運用されている読解テストのうち，英検，TOEIC，TOEFLの読解セクションにおける長文読解問題を使用した。以下に示す通り，本調査では3種類の読解テストの合計124項目の選択肢について分析を行った。

これらのテストを選んだ理由は，次のような共通した特徴を持つためである：(a)4つの選択肢から成る多肢選択式である，(b)受験者の人数が多く，大規模に実施されている，(c)正答選択肢と対応するテキスト記述が明確である。

本調査ではパラフレーズ質問の選択肢の傾向について，テスト間で比較・検討することが目的であるため，調査対象とする問題は，テキストの明示的な内容を問うパラフレーズ質問（清水，2005）に絞った。そのため，(c)を満たしていない，すなわち正答選択肢の情報についてテキスト中に対応する記述が存在していない問題に関しては，以下の分析対象とした項目の中からも除外している。加えて，受験者に対し，テキストで述べられている内容と合致していない選択肢を選ばせる問題や，テキスト全体の主旨を問う問題などは，本研究の焦点とするパラフレーズ質問とは異なるため，分析から除外した。

a. 英検2級（項目数41）（日本英語検定協会，2017; 旺文社，2017）
2015年度第1回4B, 4C, 2015年度第2回4B, 4C, 2015年度第3回4B, 4C,
2016年度第1回3B, 3C, 2016年度第2回3B, 3C, 2016年度第3回3B, 3C

b. TOEIC（項目数43）（国際ビジネスコミュニケーション協会TOEIC運営委員会，2016, 2017）
TOEIC Listening & Reading 問題集1，2よりPart 7 読解問題

c. TOEFL（項目数40）（Educational Testing Service, 2013）
TOEFL iBT リーディング演習問題1 ～ 3,実戦問題1，2

4.3 手順

2.2節で示したKeck(2006)の分類に基づき，表3のような選択肢の分類基準を作成した。選択肢のパラフレージング・レベルは，「選択肢の記述と英文両方に含まれる語の数／選択肢の記述の総語数」によって算出される割合であり，客観的な数値であったため，全ての項目について調査者1名が評定を行った。

各テストのパラフレーズ質問となる項目の正答・誤答選択肢に，本文中の対応する記述が含まれる割合に応じてパラフレージング・レベルを評定した。具体的には，Keck(2006)に基づき，以下のlinkとなる語が英文中に含まれている数をカウントし，それらの総語数に対する割合を算出した(例1)。選択肢と本文中の語との対応について，調査対象とした語は主に内容語であり，原形の変化形も含め，テキスト内の合計数を算出した。関係詞，前置詞，数字以外の冠詞については，英文全体で選択肢の記述内容と関連しない部分でも多く見られることから，本調査の対象からは除外した。

- unique links(UL)：本文において，局所的(1～2度)に表れる内容語(e.g., 名詞，形容詞，動詞)。
- general links(GL)：本文において，広範囲(3度以上)に多く表れる内容語

■表3: 選択肢のパラフレージング・レベル分類基準

	UL, GLの語数の合計が総語数に占める割合
Near Copy (NC)	50%以上
Minimal Revision (MIR)	20～49%
Moderate Revision (MOR)	1～19%
Substantial Revision (SR)	0%

■例1: 選択肢の評定方法［英検2級2016年度第3回3C (34)］

(34) What are chocolate manufacturers concerned about?
1. **People** prefer **living** a healthy life to eating sweets such as chocolate. (12語)

UL（太字）：people(1語) ＋ living (1語) ＝ 2語 → 2／12 ＝ 16.7%
GL（下線）：chocolate(1語) × 1 ＝ 1語 → 1／12 ＝ 8.3%
合計：UL(16.67%) ＋ GL(8.33%) ＝ 25.0%
⇒ Minimal Revisionと評定。

4.4 結果と考察

各テストのパラフレーズ質問となる項目の正答選択肢および錯乱肢に，本文中の対応する記述が含まれる割合に応じ，パラフレージング・レベルを評定した結果について，それぞれ表4，5に示す。

分類の結果，対象とした読解テストの正答・誤答選択肢の全体的な傾向としてはMIRが最も多く，約半数を占めており，MIRに次いで多かったのはNCであることが明らかになった。つまり，現行の読解テストにおいては，正答・誤答を問わず，ある程度英文中で使用される語を含む選択肢が多く用いられることがわかった。一方で，MOR, SRなどパラフレージング・レベルが低い選択肢の割合に関してはテスト間の違いが見られた。以下より，各テストの選択肢の特徴について，理由とともに考察し，テスト間の比較を行う。

■表4: パラフレーズ質問の正答選択肢の分類結果

	選択肢数	NC	MIR	MOR	SR
英検	41	13 (31.7%)	26 (63.4%)	2 (4.9%)	0 (0%)
TOEIC	43	10 (23.3%)	17 (39.5%)	8 (18.6%)	8 (18.6%)
TOEFL	40	13 (32.5%)	27 (67.5%)	0 (0%)	0 (0%)
Total	124	28 (28.6%)	52 (53.1%)	10 (10.2%)	8 (8.2%)

■表5: パラフレーズ質問の錯乱肢の分類結果

	選択肢数	NC	MIR	MOR	SR
英検	123	29 (23.6%)	86 (69.9%)	8 (6.5%)	0 (0%)
TOEIC	129	27 (20.9%)	34 (26.4%)	14 (10.9%)	54 (41.9%)
TOEFL	120	56 (46.7%)	58 (48.3%)	2 (1.7%)	4 (3.3%)
Total	372	112 (30.1%)	178 (47.8%)	24 (6.5%)	58 (15.6%)

4.4.1 英検

英検では全体的に，節および文単位で提示される選択肢が多かった。正答選択肢・錯乱肢ともに，MIRが最も多く，次いでNC，MORの順に多いという特徴が共通していた。そして，今回対象とした項目に限って言えば，3つのテストの中で唯一，SRの正答選択肢・錯乱肢は見られなかった。表6は英検の問題からの抜粋である。

■表6: 英検で出題された項目の例（英検2級2016年度第1回3B第1パラグラフより）

Many large snakes eat as rarely as once every three months. When they do eat, however, they eat whole animals, such as deer or wild pigs. In order to digest these large meals, the heart and other organs inside the snake's body suddenly grow much larger. The heart, for example, does this so that it can pump blood around the body much faster. This allows the snake to digest food more quickly. After the snake has digested the meal, the organs go back to their usual size.

(30) Many large snakes change the size of their hearts to in order to

1 make more space for the whole animals that they eat. (MIR)

2 better digest the blood of the animals that they eat. (MIR)

3 be able to catch more food more quickly. (NC)

④ help them digest large meals more quickly. (NC)

この設問で正答を得るためには，下線部の3文を理解することが求められる。特に，第3文を理解できれば，記述の内容およびdigest，large，mealsなどの使用されている語彙が本文と一致している正答選択肢4を選ぶことができる。

残りの錯乱肢について詳しく見てみると，いずれの錯乱肢についても，eat，animalsなど，テキスト中に複数回現れる単語や，food，more quickly，など，特定の部分において局所的に出てくる単語が含まれている。このように，本文の記述と正答選択肢・錯乱肢両方の間で語のオーバーラップの程度が高いことが分かる。

英検の項目の多くを占めていたのは，このようにNC，MIRのパラフレージング・レベルに分類される正答選択肢・錯乱肢であり，MOR，SRのレベルに該当する選択肢の数は少なかった。このような結果となった理由は，(a)選択肢の提示単位，(b)テキストのレベル，に関係していると考えられる。(a)に関しては，この後に述べるTOEICに比べ，英検の選択肢は節もしくは文単位で提示される場合が多く，記述の語数が多かったために，本文中との語のオーバーラップの割合も増加したという可能性が考えられる。(b)については，本調査では，他のテストとの比較および調査2で参加する協力者の読解熟達度を考慮したため，英検の中でも2級の問題のみをマテリアルとして扱った。例えば，準1級や1級で使用されるテキスト・選択肢を用いて分析した場合，テキストはより長くなるが設問数は減少し，1つの設問で問われる情報量は級が上がれば上がるほど多くなると考えられるため，結果として選択肢では本文で記述された複数の情報を統合した表現が用いられ，MOR，SRの選択肢が増加する可能性は十分考えられる。

4.4.2 TOEIC・TOEFL

以下では，TOEICとTOEFLの概観について併せて見ていく。

まず，TOEICでは全体として，語や句などの単位で提示される選択肢が多かった。今回扱った

■表7: TOEICで出題された項目の例(TOEIC Listening & Reading 問題集1 Part1より)

To: All Employees
From: Su Kyung Cho <skcho@greatcars.co.nz>
Subject: Annual outing
Date: 1 December

I wanted to let everyone know that management has decided to postpone our annual company hiking trip, which is usually held in December. As all of you know, that is the busiest time of year for our car rental agency, and every year it is difficult to fit the trip into our schedule. The new date is 23 January. This should be more convenient for anyone who wants to go.

(161) What is the purpose of the e-mail?

a	To schedule a meeting	(MIR)
b	To make hotel reservations	(SR)
c	To explain a car rental policy	(MIR)
ⓓ	To announce a change in plans	(MOR)

問題では，正答選択肢はMIR，NC，MORおよびSRの順で多いのに対し，錯乱肢はSR，MIR，NC，MORの順で多かった。また，他のテストに比べ，SRの選択肢が多いことが特徴的であった。表7はTOEICの問題からの抜粋である。

この問題で問われているのは書き手がメールを送った目的であり，第一文および下線部の情報に含まれる，“to let everyone know”を“to announce”へとパラフレーズすることにより，正答選択肢dを解答することができる。一方で，錯乱肢に関しては，aとcはテキスト内にそれぞれschedule，car rentalといったテキスト内で局所的に出現する語が含まれているのに対し，bに関してはいずれの語も本文中で使用されていない。また，いずれの錯乱肢においてもその内容はテキストの記述と合致していない。

このように，TOEICでは，正答選択肢ではテキストの対応する情報についてテキストで使用されている語よりも短い表現で言い換える場合が，錯乱肢においてはテキスト内で局所的に使用されている語を含んでいるか，全く対応する記述が存在しないような場合が多く見られた。

この理由には，他のテストと比較して，TOEICでは選択肢の提示単位が語・句と比較的短く，また，設問ごとのテキストの長さが短いことが挙げられる。テキストの総語数が少なく，選択肢の提示単位のために本文よりも短い表現で言い換える必要性から，テキスト中と重複する語が限定される可能性が考えられる。また，そうして作成される錯乱肢の数には限りがあるため，テキスト内で対応する記述が全く存在しない錯乱肢も含めて提示している可能性もある。

一方，TOEFLにおいては，句単位あるいは節・文単位で提示される選択肢が多く見られており，正答選択肢はMIR，NCのみに集中し，錯乱肢はMIR，NC，SR，MORの順で多かった。表8はTOEFLの問題からの抜粋である。

この設問の正答選択肢cは，テキスト本文中の下線部の情報を理解することで解答でき，受験者には，本文中の“had almost no contact with

■表8: TOEFLで出題された項目の例(TOEFL iBT リーディング演習問題2, The Expression of Emotionsより)

Moreover, people in diverse cultures recognize the emotions manifested by the facial expressions. In classic research Paul Ekman took photographs of people exhibiting the emotions of anger, disgust, fear, happiness, and sadness. He then asked people around the world to indicate what emotions were being depicted in them. Those queried ranged from European college students to members of the Fore, a tribe that dwells in the New Guinea highlands. All groups, including the Fore, who had almost no contact with Western culture, agreed on the portrayed emotions.
The Fore also displayed familiar facial expressions when asked how they would respond if they were the asked in stories that called for basic emotional responses. Ekman and his colleagues more recently obtained similar results in a study of ten cultures in which participants were permitted to report that multiple emotions were shown by facial expressions. The participants generally agreed on which two emotions were being shown and which emotion was more intense.

(5) According to paragraph 2, which of the following was true of the Fore people of New Guinea?

a They did not want to be shown photographs. (MIR)

b They were famous for their story-telling skills. (SR)

ⓒ They knew very little about Western culture. (MIR)

d They did not encourage the expression of emotions. (MIR)

Western culture”という記述が選択肢において“knew very little about Western culture”へと言い換えられていることに気づくことが求められている。一方，錯乱肢に関しては，a，dについては局所的（e.g.，shown，photographs）および全体的に（e.g.，expression，emotions），本文中でも出現する語が使われているものの，内容が対応する記述は存在しない。また，錯乱肢bに含まれる語および情報も，本文には記述されていない。

上記のようにTOEFLにおいては，正答にはテキストの対応する語を比較的高い割合で含む選択肢が多く，錯乱肢においてはテキスト内で局所的又は全体的に使用されている語を含むか，全く対応する記述が存在しないような場合が見られた。この理由として，TOEFLでは他のテストに比べ，1設問当たりに関連するテキストの分量が多いという特徴が考えられる。このように総語数，およびパラグラフごとの分量も多いテキストでは，設問ごとに問われる範囲が広く，選択肢作成における自由度が比較的高くなり，NC・MIRの選択肢を量的に多く作ることや，多様なレベルでの錯乱肢の作成がしやすかったと考えられる。

4.4.3　調査1のまとめ

3種類の多肢選択式テストのパラフレーズ質問における選択肢を比較した結果，正答・誤答選択肢のパラフレージング・レベルの全体的な傾向としてはMIRが最も多く，約半数を占めており，次いで多かったのはNCであることが明らかになった。つまり，現行の読解テストにおいてパラフレーズ質問で求められている能力は，正答・誤答を問わず，ある程度英文中で使用される語を含む選択肢を選ぶことのできる能力であると考えられる。

一方で，MOR，SRなどパラフレージング・レベルが低い選択肢の割合に関してはテスト間の違いがあることがわかった。今回調査した読解問題において，英検ではSRの選択肢は正答・誤答いずれも見られず，TOEFLではMOR，SRの正答選択肢は見られなかったが，TOEICでは他の2つのテストに比べてSR，MORの正答・誤答選択肢が多くなっていた。

上記の分類結果を踏まえ，調査2では，既存の読解テストにおけるパラフレーズ質問の正答選択肢および錯乱肢のパラフレージング・レベルを改編した読解テストを作成する。テキスト中の語とのオーバーラップの割合が異なる選択肢によって構成される読解問題を，日本人英語学習者である大学生を対象に実施することにより，各選択肢の受験者を引き付けている程度がどのように異なるのかを検証する。

5　調査2

5.1　目的

調査1より，現行の読解テストにおけるパラフレーズ質問の選択肢のパラフレージング・レベルの傾向はテスト間で異なることがわかった。そこで調査2では，調査1の分類手順に基づいて既存の読解テストの選択肢を改編し，4つの異なるパラフレージング・レベルの選択肢からなる読解問題を作成・提示し，受験者が各選択肢を選んだ割合を明らかにする。加えて，各選択肢が受験者にとってどの程度もっともらしいのかを示すもう一つの指標として，魅力度を明らかにする。また，本調査では，テストの総得点に応じて受験者を上位群，中位群，下位群の3つに分類し，受験者のパフォーマンスによって，各選択肢の選択率や魅力度がどのように異なるのかについても検討する。

5.2　協力者

国立大学に通う43名の日本人大学生・大学院生が調査に参加した。ただし，協力者のうち，制限時間内に全ての項目に解答できなかった1名のデータは分析から除外した。協力者の専攻は，医学，工学，生物学，人文学，社会学，教育学など多岐にわたった。

5.3　マテリアル

5.3.1　使用したテスト

調査2では，英検2級過去問題の長文読解セクションより選定した以下の4つの読解問題を，基盤のテストとして用いることとした（表9）。その

上で，基盤としたテストの各項目の正答選択肢および錯乱肢について，4つのパラフレージング・レベルが同じ数ずつ設けられるように改変したテストをマテリアルとして使用した。

英検の読解問題を基盤のテストとした理由は，(a)選択肢の記述の形態が節・文で統一されていること，(b)本文中で使用される語を含んでいる割合が低い選択肢の出現頻度が少ないこと，の2点が挙げられる。

(a)に関しては，調査1で明らかになったように，英検の選択肢は節あるいは文の形で記述されている。本調査においては，テキスト原文との語のオーバーラップの割合を調整することで選択肢のパラフレージング・レベルを設定するため，記述のおおよその語数や提示単位など，各項目の選択肢間の等質性が確保されている点で，英検の読解問題が最適であると考える。

また，(b)については，調査1で明らかになったこととして，現行の読解テストの中でも英検は，本文中で使用される語を含んでいる割合が低い選択肢を用いた項目が少ないことが挙げられる。特に，SRに関しては正答選択肢・錯乱肢のいずれにおいても例が見られなかった。ゆえに，正答選択肢と錯乱肢のパラフレージング・レベルの組み合わせによる受験者への影響を調べる上で，テキスト原文との重複の程度が低い選択肢を用いた検証の余地がより大きいと考えた。

■表9: 調査2で使用した英検過去問より選んだ英文 (旺文社, 2003)

テキスト	総語数	FKGL
2001年度第2回4A　Green Cleaners	338	8.4
2001年度第2回4B　Drugstore Dilemma	320	10.3
2002年度第1回4B　The Placebo Effect	338	11.3
2001年度第3回4B　Million-Dollar Math	342	10.7

(注) FKGL = Flesch-Kincaid Grade Level.

5.3.2　錯乱肢の作成

初めに，基盤とした英検の読解問題について，4.3節例1に示す通りの手順で，選択肢中にテキスト原文の語が含まれる割合を算出した。続いて，4択問題における正答・誤答の選択肢のパラフレージング・レベルについて，正答選択肢のレベルが決まると他の錯乱肢は残りの3つのレベルになるような問題パターンを4つ設定した(表10)。

1つの問題パターンごとに4つの選択肢にパラフレージング・レベルを振り分けられるよう，算出して得られたテキスト原文の語の割合と，錯乱肢のパターン内におけるレベルが異なる場合には調整を行った。具体的には，当該選択肢が設定したいパラフレージング・レベルよりも高い場合には，テキスト原文の記述をパラフレーズした表現(e.g.，上位語，類義語)を使用し，逆に低い場合には，テキスト原文で用いられている表現を使用することにより，修正を行った。調査者が改編した選択肢は，ネイティブスピーカーによって表現や意味に誤りや不自然さがないかを確認され，必要な場合には修正が加えられた。

■表10: 調査2で作成したマテリアルの問題パターン

	P 1	P 2	P 3	P 4
正答選択肢	NC	MIR	MOR	SR
錯乱肢①	MIR	MOR	SR	NC
錯乱肢②	MOR	SR	NC	MIR
錯乱肢③	SR	NC	MIR	MOR

(注) P = パターン (pattern)。

5.3.3 自信度アンケート

協力者の正答選択肢および錯乱肢に対する捉え方を調べるため，読解テストの解答用紙の解答記入欄の隣に，各選択肢への解答の自信度を回答する質問項目を設けた（資料2：110頁参照）。協力者は各選択肢の正誤の解答について，低・中・高の3段階のいずれかで自信度を答えるよう指示され，該当する自信度に丸をつけて回答した。

5.3.4 筆記再生課題

筆記再生課題を使用する目的は，本調査で使用した読解テストの項目は全てパラフレーズ質問であり，表層的記憶および命題的テキストベースの理解を測定していると考えられるが，同様のテキスト理解を測定していると考えられる筆記再生課題について，その成績と読解テストの得点との相関を調べることにより読解テストの妥当性を確認するためであった。筆記再生課題は紙面にて作成・実施し，協力者は，各英文に関して覚えている内容をできるだけ多く書き出すよう指示され，記述する際は単語の羅列でなく，文章の形で書くことを求められた。本調査で使用したテキストの数は4つと多かったため，協力者は課題に解答する際に，内容を想起する手がかりとして各英文のタイトルを与えられた。

5.4 手順

調査は個別，もしくは4名程度をまとめて一斉に実施された。初めに実験の目的と，個人情報は特定不可能な形で厳重に管理されることを伝え，実験参加への承諾を得た。続いて実験全体の流れと，リーディングテストの問題解答，選択肢の自信度回答の方法について，例を提示しながら説明を行った。

説明の後，協力者は2つの英文についての読解テスト（20分），筆記再生課題（時間無制限）の順に紙面上で解答を行い，5分間の休憩の後に同様の手順を繰り返し，合計4つの英文について，読解テストと筆記再生に取り組んだ。読解する4つの英文の順番は，カウンターバランスをとった。また，通常のテスト実施時と同様，設問解答の際にはテキストを参照することを可能とした。調査はほとんどの場合，90分程度で終了した。

5.5 採点・分析

5.5.1 リーディングテスト

多肢選択式問題については，正答1項目につき1点を与え，合計得点を算出した。また，各選択肢が選ばれた割合を算出した。

分析について，本調査では有意水準を $\alpha = .05$ として設定し，$\alpha < .10$ の場合に有意傾向とみなすこととする。球面性が仮定されない要因についてはGreenhouse Geisser法による修正を採用し，多重比較についてはBonferroni法による修正を行った。

5.5.2 自信度アンケート

回答データは，受験者の正誤判断と自身の解答に対する自信度に応じ，1〜6の6段階でコーディングされた（表11）。これ以降では，正誤判断と解答の自信度を組み合わせることで得られた評定値を「魅力度」として扱う。

■表11：正答選択肢・錯乱肢の魅力度の算出方法

回答自信度	正誤判断	魅力度
高	正	6
中	正	5
低	正	4
低	誤	3
中	誤	2
高	誤	1

5.5.3 筆記再生課題

Ikeno（1996）に基づき，英語教育学を専攻する大学院生2名と調査者により4つの英文をアイディアユニット（IU）に分割した（一致率：91.90%、93.46%）。採点では各IUにつき，約3分の2の情報が再生された場合に1点を与えた。初めに解答データ30%について，大学院生2名と調査者がペアで採点し（一致率：96.57%、90.18%），不一致点を協議で解決した後，さらに明確化した基準に基づき，残りのデータは調査者1名が採点した。

5.6 結果

5.6.1 リーディングテストの信頼性

本調査で実施した全20項目のテストについて信頼性係数を算出したところ，α = .69であった。そのため，信頼性が低い4項目を除外したところ，16項目でα = .72の値が得られた。本調査の項目数およびサンプルサイズは小さいため，ある程度の信頼性が確保されたといえる。表12は，項目を修正した後のリーディングテストの記述統計および内的一貫性を示している。

■表12: リーディングテストの記述統計と信頼性係数 (K=16)

N	*M*	*SD*	*Max*	*Min*	α
42	12.60	2.71	16	5	.72

（注）最高点は15点。

5.6.2 リーディングテストの妥当性

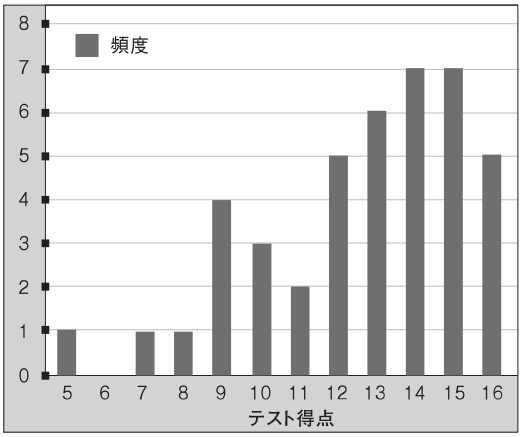
本調査のマテリアルは，英検で使用されたテスト項目を用いたため，基本的にはテストとしての妥当性は満たされていると考えられる。ただし，留意すべき点として，元の選択肢において含まれていた語を改編して作成した項目であったことが挙げられる。

Messick (1995) はテストの構成概念妥当性に関し，その中において，(a) 内容的側面，(b) 本質的側面，(c) 構造的側面，(d) 一般化可能性の側面，(e) 外的側面，(f) 結果的側面の6つの側面があると規定している。本調査では，この中の (b) 本質的側面に関して，使用したテスト項目の妥当性検証を行った。具体的には，各協力者の多肢選択式読解テストの得点と筆記再生課題で得られた再生率との相関を分析し，テスト解答時に達成されているテキスト理解から，読解後の記憶表象に保持されている内容の程度は予測できるかどうかを検証した。

筆記再生率のデータが正規分布を満たさないことから，ノンパラメトリック検定であるスピアマンの順位相関係数 (ρ) を用いて分析したところ，ρ = .783 (p = .003) という強い相関が示された。これより，本調査で使用した読解テストは，英文の表層的記憶・命題的テキストベースの理解を測定しているというある程度の妥当性が確認された。

5.6.3 テスト得点

分析対象とした16項目について，協力者の平均得点，標準偏差を表13に示す。全協力者42名の平均値は12.60点であった。今回のテスト得点別の人数分布の傾向（図1），および，上位群27%，下位群27%，中位群46%に分けることが一般的な言語テスト分野の慣習に基づき (Hughes, 2003)，協力者のうち11点以下の者を下位群 (n = 12)，12～14点の者を中位群 (n = 18)，15, 16点の者を上位群 (n = 12) に分類した。一元配置分散分析により，上中下群間の総得点には有意な差があることが確認された，$F(2, 39) = 106.67$, $p < .001$, $\eta^2 = .85$。



■図1: テスト得点別の人数分布

次節からは，協力者全体，および上記のようにして得られたテスト得点上位・中位・下位の3つの群の傾向について詳細に調査する。具体的には，各パラフレージング・レベルの選択肢の選択率と魅力度それぞれについて，全体的傾向を概観した後，テストの得点群ごとの傾向について調査する。

■表13: テスト得点群別の協力者の得点

	N	*M*	*SD*	*Min*	*Max*
上位	12	15.42	0.51	15.00	16.00
中位	18	13.11	0.83	12.00	14.00
下位	12	9.00	1.71	5.00	11.00
合計	42	12.60	2.71	5.00	16.00

5.6.4 選択肢の選択率

表14に各選択肢が選ばれた割合を示す。まず，これらの4つの選択肢の選択率について，一元配置分散分析によって比較したところ，NCの選択率がSRよりも有意に高くなっていた（p = .007）。一方，MIR，MORの選択率については，他の種類との差は確認されなかった（ps > .05）。これらのことから，全体的な傾向として，選択肢の記述内に本文で使用される語が50%以上含まれる場合，本文で使用される語をほとんど含んでいない選択肢よりも受験者に選ばれる可能性が高いことがわかった。

■表14: 各選択肢が選ばれた割合

NC		MIR		MOR		SR	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
0.27	0.07	0.26	0.08	0.26	0.08	0.22	0.06

続いて，このような各選択肢の選択率の傾向が，協力者のテスト得点および正答の選択肢の種類によってどの程度異なるのかを検証する。

4つの問題パターンにおける，テスト得点群別の各選択肢の選択率を表15および図2に示す。協力者ごとの各選択肢の選択率を従属変数とし，3（総得点：上位，中位，下位）× 4（問題パターン：P1，P2，P3，P4）×4（選択肢の種類：NC，MIR，MOR，SR）の三元配置分散分析を行った。その結果，選択肢の種類の主効果，$F(3, 117) = 3.46$，p = .019，$\eta^2 < .01$，および，問題パターン×選択肢の種類の交互作用，$F(6.68, 260.55) = 321.50$，p < .001，$\eta^2 = .79$，問題パターン×選択肢の種類×総得点の交互作用，$F(13.36, 260.55) = 15.36$，p < .001，$\eta^2 = .08$がそれぞれ有意であった。

以下では，有意であった，問題パターン×選択肢の種類×総得点の交互作用の解釈を行う。この交互作用についてさらに検証するために下位検定を行った結果，総得点の単純・単純主効果が，3つの組み合わせ（P1: SR，P2: NC，P4: MOR）を除いた13個全ての問題パターン×選択肢の種類の組み合わせにおいて有意であった（ps < .05）。つまり，正答選択肢のパラフレージング・レベルと受験者の得点により，各選択肢の選ばれる割合が異なっていたことが示された。

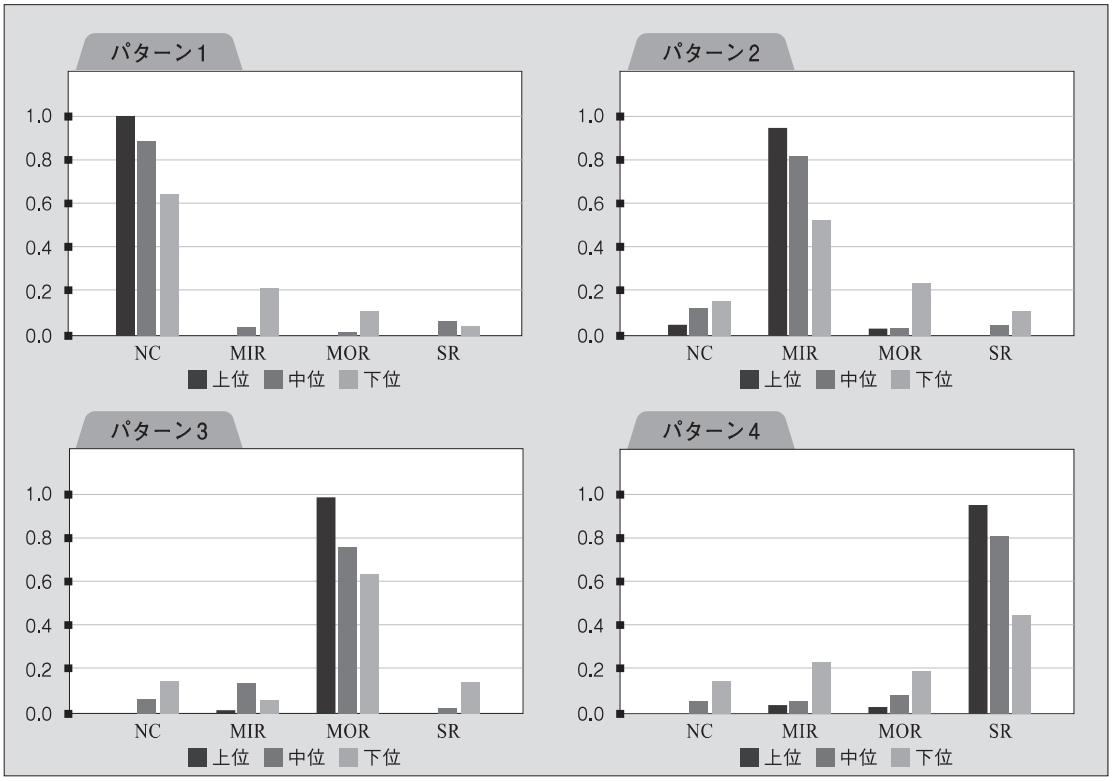
各問題パターン（P）において選択肢が選ばれた割合について，得点群間で多重比較を行った結果，下位群はP1，P2，P4においてそれぞれ正答のレベル（NC，MIR，SR）を選んでいる割合が上位群，中位群よりも低く（ps < .001），P3において正答のMORを選んでいる割合が上位群よりも低かった（p < .001）。錯乱肢に関しては，P1においてMIR，MOR（ps < .05），P2においてMOR（ps < .001），P3においてSR（ps < .001），P4においてMIR（ps < .01）を選んでいる割合が，それぞれ上位群・中位群よりも高いことが示された。さらに，下位群は上位群よりも，P3においてNC（p = .004），P4においてNC（p = .002）を選んでいる割合が有意に高かった。

これらの結果をまとめると，正答選択肢がいずれのパラフレージング・レベルであっても，下位群は上位群・中位群よりも正答を選ぶことが困難であり，他のレベルで作成された錯乱肢をより多く選択していたことがわかった。また，上位

■表15: 各問題パターンにおける各選択肢の選択率

問題	熟達度	NC		MIR		MOR		SR	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
P1 (NC)	上位	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	中位	0.89	0.18	0.04	0.10	0.01	0.06	0.06	0.11
	下位	0.65	0.17	0.21	0.18	0.10	0.13	0.04	0.10
	合計	0.85	0.20	0.08	0.14	0.04	0.09	0.04	0.09
P2 (MIR)	上位	0.04	0.10	0.94	0.11	0.02	0.07	0.00	0.00
	中位	0.11	0.15	0.82	0.21	0.03	0.08	0.04	0.10
	下位	0.15	0.13	0.52	0.20	0.23	0.20	0.10	0.13
	合計	0.10	0.14	0.77	0.24	0.08	0.15	0.05	0.10
P3 (MOR)	上位	0.00	0.00	0.02	0.07	0.98	0.07	0.00	0.00
	中位	0.07	0.12	0.14	0.13	0.76	0.18	0.03	0.08
	下位	0.15	0.13	0.06	0.11	0.65	0.25	0.15	0.13
	合計	0.07	0.11	0.08	0.12	0.79	0.22	0.05	0.10
P4 (SR)	上位	0.00	0.00	0.04	0.10	0.02	0.07	0.94	0.11
	中位	0.06	0.11	0.06	0.11	0.08	0.15	0.81	0.16
	下位	0.15	0.13	0.23	0.20	0.19	0.24	0.44	0.19
	合計	0.07	0.11	0.10	0.16	0.10	0.17	0.74	0.25

（注）P = パターン (pattern) 。(　)内は正答選択肢の種類を示す。



■図2: 各問題パターンにおける各選択肢の選択率

群に比べ，MOR，SRといった正答選択肢のパラフレージング・レベルが低い場合に，テキストと重複する語の割合が最も高いNCの錯乱肢をより多く選択する傾向があったことが示された。

一方，中位群と上位群の選択率の間に差が見られた項目は，P3であった。P3において，上位群は中位群よりも正答のMORを選んでいる割合が高かったのに対し（p = .009），中位群は上位群に比べ，MIRをより多く選ぶ傾向があったことがわかった（p = .020）。すなわち，正答選択肢のパラフレージング・レベルがMORであったとき，上位群に比べ，中位群の受験者はそれを選ばず，テキスト原文との重複している語がより多く含まれるMIRを選ぶ傾向にあったということが示唆された。

5.6.5 選択肢の魅力度

表16は，各選択肢の魅力度の評定値を示している。まず，全体的な4つの選択肢の魅力度について，一元配置分散分析で比較した結果，NCの評定値がMOR（p = .002），SR（p = .006）よりも高いことがわかった。MIRについては他の種類との差は見られなかった（ps > .05）。これらのことから，全体的な傾向として，選択肢の記述内に本文で使用される語が50%以上含まれる場合，本文で使用される語を含む割合が20%程度以下の選択肢よりも受験者を引き付けやすく，正答と判断される確率が高いことがわかった。

次に，各パラフレージング・レベルで作成された正答選択肢および錯乱肢の魅力度について，協力者のテスト得点群ごとに評定値の傾向は異なるのかどうかを検証するため，各テスト得点群の4つの種類の正答選択肢・錯乱肢に対する魅力度評定値を従属変数とし，3（総得点：上位，中位，下位）×4（選択肢の種類：NC，MIR，MOR，SR）の二元配置分散部分析を行った。

まず，正答選択肢に関しては，選択肢の種類の主効果が有意であり，$F(3, 117) = 6.58$, $p < .001$, $\eta^2 = .07$, NCがMORよりも高く（$p < .001$），SRよりも高い傾向にあることが明らかになった（$p = .026$）。一方で，総得点の主効果，$F(2, 39) = 0.18$, $p = .084$, $\eta^2 < .01$，および，選択肢の種類×総得点の交互作用$F(6, 117) = 0.85$, $p = .534$, $\eta^2 = .02$，は有意でなかった。つまり，テス

■表16: 各選択肢の魅力度

問題	熟達度	全体		正答		誤答	
		M	*SD*	*M*	*SD*	*M*	*SD*
NC	上位	2.44	0.18	2.40	0.16	0.92	2.15
	中位	2.65	0.30	2.49	0.25	3.42	2.56
	下位	2.64	0.45	2.34	0.45	4.74	1.58
	合計	2.58	0.33	2.42	0.30	3.08	2.62
MIR	上位	2.39	0.32	2.34	0.31	1.25	2.30
	中位	2.52	0.39	2.36	0.33	3.69	2.40
	下位	2.69	0.52	2.37	0.43	4.60	1.55
	合計	2.53	0.42	2.36	0.35	3.26	2.50
MOR	上位	2.24	0.24	2.21	0.23	0.83	1.99
	中位	2.23	0.30	2.15	0.26	1.89	2.47
	下位	2.51	0.41	2.20	0.39	4.42	1.49
	合計	2.31	0.34	2.18	0.29	2.31	2.49
SR	上位	2.34	0.22	2.34	0.22	0.00	0.00
	中位	2.33	0.36	2.26	0.34	1.86	2.41
	下位	2.38	0.52	2.17	0.44	4.25	2.09
	合計	2.35	0.37	2.26	0.34	2.01	2.49

トの得点群に関わらず，正答選択肢においてテキスト原文と同じ語が使用されている割合が高い場合，割合が低い場合よりも，受験者に自信をもって選ばれやすかったことがわかった。

一方，錯乱肢に関しては，総得点の主効果が有意であり，$F(2, 39) = 51.62$, $p < .001$, $\eta^2 = .31$, 魅力度は下位群において上位群（$p < .001$），中位群（$p < .001$）よりも高く評定され，さらに，中位群において上位群よりも高く評定されていた（$p < .001$）。他の，選択肢の種類の主効果，$F(3, 117) = 2.44$, $p = .068$, $\eta^2 = .03$, および，選択肢の種類×総得点の交互作用$F(6, 117) = 0.63$, $p = .705$, $\eta^2 = .02$, は有意でなかった。このことから，錯乱肢の記述においてテキスト原文と同じ語を含む割合の高低に関わらず，総得点の高い受験者に比べ，低い受験者ほど，錯乱肢に引き付けられやすかったことが示唆された。

続いて，このような選択肢の魅力度について，選択率と同様に，協力者のテスト得点および正答の選択肢の種類によってどの程度異なるのかを検証する。4つの問題パターンにおける，テスト得点群別の各選択肢の魅力度を表17および図3に示す。協力者ごとの各選択肢の魅力度を従属変数とし，3（総得点：上位，中位，下位）× 4（問題パターン：P1, P2, P3, P4）× 4（選択肢の種類：NC, MIR, MOR, SR）の三元配置分散分析を行った。その結果，問題パターンの主効果，$F(3, 117) = 7.10$, $p < .001$, $\eta^2 < .01$, および，問題パターン×選択肢の種類の交互作用，$F(6.11, 238.25) = 307.14$, $p < .001$, $\eta^2 = .76$, 問題パターン×選択肢の種類×総得点の交互作用，$F(12.22, 238.25) = 12.49$, $p < .001$, $\eta^2 = .06$がそれぞれ有意であった。

ここからは，有意であった，問題パターン×選択肢の種類×総得点の交互作用の解釈を行う。この交互作用についてさらに検証するために下位検定を行った結果，総得点の単純・単純主効果が，5つの組み合わせ（P1: MOR, SR, P2: NC, SR, P3: MIR）を除き，11個全ての問題パターン×選択肢の種類の組み合わせにおいて有意であった（$ps < .05$）。つまり，正答選択肢のパラフレージング・レベルと受験者の得点により，各選択肢の魅力度が異なっていたことがわかった。

各問題パターンにおける選択肢の魅力度についてテスト得点群間で多重比較を行った結果，P1, P2, P4の正答の選択肢（NC, MIR, SR）の下位群に対する魅力度は上位群，中位群よりも低く（$ps < .01$），また，P3の正答MORの下位群に対する魅力度は上位群よりも低かった（$p = .005$）。一方，錯乱肢に関しては，P1のMIR（$ps < .05$），P2においてMIR, MOR（$ps < .01$），P4のMIR（$ps < .05$）の下位群に対する魅力度が，それぞれ上位群・中位群よりも高いことが示された。さらに，下位群は上位群に比べ，P3においてNC（$p = .004$），P4においてNC, MOR（$ps < .01$）の魅力度評定値が有意に高かった。

これらの結果をまとめると，どのパラフレージング・レベルで作成された正答選択肢であっても，下位群は上位群・中位群よりも強い自信をもって正答と判断することが困難であり，他のレベルで作成された錯乱肢により引き付けられていたことがわかった。また，上位群に比べ，MOR, SRといった正答選択肢のパラフレージング・レベルが低い場合に，テキストと重複する語の割合がより高いNC, MORの錯乱肢をより魅力的に感じる傾向があったことが示された。

一方，中位群と上位群の差に関しては，P3におけるMOR, MIRで差が見られた選択率の結果とは異なり，いずれの項目においても中位群と上位群の魅力度評定値の間に統計的有意差は確認されなかった。

5.7 結果のまとめと考察

本調査において得られた主な結果は，以下の4点にまとめられる。

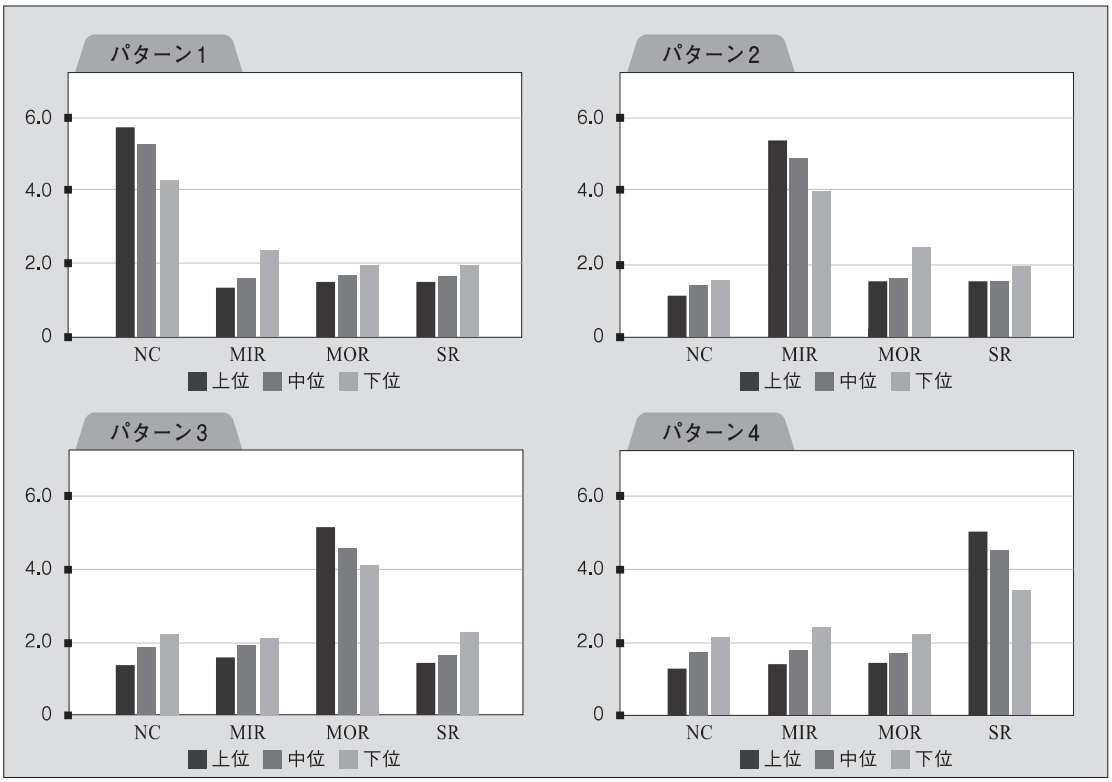
a 全体的に，テキスト原文と同じ語が50%以上含まれる選択肢は，本文で使用される語を0〜20%程度しか含まない選択肢よりも受験者に選ばれやすく，自信をもって正答だと判断されやすい。

b 錯乱肢の記述においてテキスト原文と同じ語を含む割合の高低に関わらず，高得点の受験者よりも低得点の受験者ほど，錯乱肢を魅力的であると判断する。

■表17: 各問題パターンにおける選択肢の魅力度

問題	熟達度	NC		MIR		MOR		SR	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
P1 (NC)	上位	5.67	0.34	1.31	0.36	1.42	0.51	1.44	0.40
	中位	5.19	0.73	1.57	0.53	1.61	0.65	1.63	0.48
	下位	4.21	0.73	2.31	0.94	1.90	0.59	1.90	0.67
	合計	5.05	0.85	1.71	0.74	1.64	0.61	1.65	0.54
P2 (MIR)	上位	1.15	0.35	5.23	0.52	1.54	0.53	1.50	0.44
	中位	1.43	0.44	4.76	0.65	1.64	0.40	1.54	0.43
	下位	1.60	0.57	3.92	0.86	2.46	0.66	1.94	0.83
	合計	1.40	0.48	4.65	0.84	1.85	0.64	1.64	0.59
P3 (MOR)	上位	1.38	0.38	1.58	0.43	5.13	0.56	1.42	0.40
	中位	1.86	0.56	1.93	0.67	4.49	0.74	1.65	0.66
	下位	2.21	0.76	2.13	0.57	4.08	0.92	2.27	0.62
	合計	1.82	0.65	1.89	0.61	4.55	0.83	1.76	0.67
P4 (SR)	上位	1.27	0.27	1.42	0.39	1.44	0.40	5.02	0.54
	中位	1.74	0.60	1.81	0.53	1.72	0.54	4.50	0.67
	下位	2.15	0.55	2.42	0.65	2.25	0.88	3.40	0.82
	合計	1.72	0.60	1.87	0.65	1.79	0.69	4.33	0.92

（注）P = パターン (pattern)。( )内は正答選択肢の種類を示す。



■図3: 各問題パターンにおける各選択肢の魅力度

c 正答選択肢がテキスト原文と同じ語を含んでいる割合の高低に関わらず，低得点の受験者は高・中程度の得点の受験者よりも正答を選ぶことが困難であり，他の錯乱肢をより多く選択した。特に，高得点の受験者に比べ，正答選択肢がテキスト原文と同じ語を含んでいる割合が低い場合に，テキストと重複する語の割合がより高い錯乱肢に引き付けられやすい。

d 正答選択肢のテキスト原文と同じ語を含む割合が1～19%の場合，高得点の受験者は正答を中程度の得点の受験者よりも多く選ぶ一方，中程度の得点の受験者はテキストと重複する語の割合がより高い（20～49%）錯乱肢を選ぶ傾向がある。

まず，aに関しては，先行研究から予測された点と一致する。Ushiro et al.（2008）では，多肢選択式テストへの解答時には，受験者が選択肢から得られる情報に依存し，特にテキストと選択肢の語が重複しているものを選ぶ傾向があることが示唆されていた。本調査においても，選択肢が選ばれた割合，そして選択肢の魅力度の観点からも，受験者が選択肢とテキストの語が重複している程度が高いほど魅力的であると判断し，引き付けられる傾向にあったと考えられる。

次に，bに関しても，先行研究と同様の結果が得られた。多肢選択式リスニングテストの選択肢の魅力度を調査したIimura（2014）において，熟達度の低い受験者は高い受験者よりも錯乱肢を魅力的だと判断する傾向があることが示されていた。本調査においても同様に，読解テストにおけるパラフレーズ質問においてパラフレージング・レベルを操作した選択肢を用いた場合，低得点の受験者は高・中程度の得点の受験者よりも，中程度の得点の受験者は高得点の受験者よりもテキストの理解が十分でなかったために，錯乱肢をもっともらしく感じる度合いが高かった可能性が考えられる。

続いてcであるが，高・中程度の得点の受験者は低得点の受験者よりも正答選択肢を多く選んでいることが明らかであるため，これは当然の結果であるといえる。しかし，低得点の受験者は高得点の受験者よりもテキスト理解が十分でなかったために，テキスト原文と同じ語を含んでいる割合が低い正答選択肢を選ぶことができず，テキストと重複する語の割合がより高い錯乱肢に引き付けられる傾向にあったと考えられる。

dに関しては，本調査において唯一，高得点・中程度の得点の受験者を区別する結果となった。すなわち，中程度の得点の受験者は高得点の受験者よりもテキスト理解が十分でなかったために，テキスト原文と同じ語を含んでいる割合が低い正答選択肢を選ぶことができず，テキストと重複する語の割合がより高い錯乱肢に引き付けられたと考えられる。このことから，パラフレーズ質問の選択肢においてテキストの理解が中程度の受験者と十分に達成されている受験者を分ける閾値は，このMIR（20～49%），MOR（1～19%）というレベルである可能性が考えられる。一方で，魅力度の側面から考えると，正答をMORとするP3において，MIRに対する魅力度評定値に，中・高得点の受験者間で有意差は見られなかった。つまり，中程度の得点の受験者は自信をもって錯乱肢であったMIRを正答と判断したわけではなく，判断を迷った結果として，消極的に選択したという可能性も考えられる。

6 結論と今後の課題

本研究では2つの調査を通して，多肢選択式読解テストにおいて，パラフレージング・レベルをパラフレーズ質問の選択肢作成に利用する可能性を検証した。以下に，本研究から得られた主要な結果をまとめる。

調査1では，英検，TOEIC，TOEFLの読解問題においてパラフレーズ質問を対象に，パラフレージング・レベルの観点から，正答・誤答の選択肢が含む語と英文中の語はどの程度重複しているのかについて調査を行った。結果から，テストの種類によって一部のパラフレージング・レベルの傾向が異なることが示された。全体的な傾向としてはMIRが最も多く，約半数を占めており，次いで多かったのはNCであった一方，MOR，SRなどパラフレージング・レベルが低い選択肢の割合に関してはテスト間に違いがあることが明ら

かになった。英検やTOEFLではMOR, SRの正答選択肢の数が少ないのに対し, TOEICでは他の2つのテストに比べてMOR, SRの正答・誤答選択肢が多い傾向があることがわかった。

続く調査2では, パラフレージング・レベルにより作成された, テキスト原文で使用される語を含む割合が異なる正答選択肢・錯乱肢, およびその組み合わせが効果的に機能するかどうかについて, 選択率および選択肢の魅力度の観点から検証した。結果, 正答選択肢・錯乱肢の組み合わせによる選択率と魅力度への影響は, テスト得点群によって異なることが示された。選択率に関して, テスト得点が低い又は中程度の受験者は, 高得点の受験者よりも, 正答選択肢のパラフレージング・レベルが低い場合に, よりレベルが高い錯乱肢を多く選択することが示された。魅力度については, テスト得点が低い受験者は高得点の受験者よりも, 正答選択肢のパラフレージング・レベルが低い場合に, それより高いレベルの錯乱肢をより魅力的であると感じることが示された。

総括すると, パラフレージング・レベルの異なる錯乱肢の機能は, 正答選択肢との組み合わせ, および, 受験者の得点によって異なることが明らかになった。すなわち, 高得点の受験者にとっては, 正答選択肢および錯乱肢のパラフレージング・レベルに関わらず, 正答を選んでいた割合は同程度であった。一方, 得点が低い又は中程度の受験者の場合は, 高得点の受験者に比べ, 選択肢のテキスト原文の語を含む割合の高低に影響を受け, 特に正答選択肢のパラフレージング・レベルが低い場合に, よりレベルが高い錯乱肢を多く選ぶ割合が高かった。特に, 中程度の得点の受験者と高得点の受験者を区別する境目となったのは, 正答が原文に含まれる語を1〜19%程度含む場合に, それよりも多い割合(20〜49%)で含む錯乱肢に引き付けられるか否かであった可能性が示された。一方で, 魅力度の観点から考えると, 中程度の得点の受験者は錯乱肢であったMIRについて, 判断を迷った結果, 消極的に選択したという可能性も否定できない。このような受験者の誤った選択を導く原因を明らかにするには, 今後の研究において, オンラインの手法を併用しながら解答プロセスをより詳細に調べる必要があるだろう。これらの結果より, 弁別力のあるパラフレーズ質問の項目を作成する際には, 選択肢のパラフレージング・レベルすなわち文章中の記述を含む割合について, そして正答選択肢と錯乱肢の組み合わせ, 受験者の判断する選択肢の魅力度を合わせて考慮することが有用であることが示唆された。

これらの調査によって, 本研究はパラフレーズ質問の項目作成における, 選択肢の表層的な性質について検討する意義を見直す機会となったといえる。表層的性質はもっともらしさとの関わりが希薄である可能性を指摘する研究もあるが, パラフレーズ質問など, 質問タイプによっては難易度を左右しうる要因として考慮するべきではないかと考えられる。さらに, 本調査では, 選択肢に対する回答自信度と正誤判断を組み合わせた魅力度を導入したことで, 各選択肢に対する学習者の認知面についても詳細に調べることが可能となった。今後も多角的な手法により, 選択肢要因と学習者の認知面が統合的に調査されることが望まれる。

最後に, 本研究における限界点として, 調査2において作成した読解テストの信頼性, 妥当性, 項目数, 一般化可能性が十分ではなかったことが挙げられる。特に, 今回基盤のテストとして用いたのは英検2級の読解問題のパラフレーズ質問であった。ゆえに, 本研究での成果について, 英検の他の級においても同様の結果が得られるのかどうか, さらに, 英検の形式以外の読解テストに応用することができるかどうかについては, 今後さらなる調査が必要となるだろう。本調査ではパラフレージング・レベルに焦点を当てたが, その基準の再設定も含め, より幅広い観点から多肢選択式問題における効果的な錯乱肢の性質について検証することが望まれる。

謝 辞

本研究を実施する機会を与えてくださいました公益財団法人 日本英語検定協会の皆さま，選考委員の先生方に厚く御礼申し上げます。特に，私の研究を担当してくださいました村木英治先生には，貴重なご助言を賜りましたこと深く感謝申し上げます。

そして，筑波大学大学院の卯城祐司先生には，本研究の計画から実施，報告書の作成に至るまで親身にご指導をいただきましたこと，心より御礼申し上げます。最後に，本調査を実施するにあたり，採点作業や調査にご協力をいただいた皆様に厚く御礼申し上げます。

参考文献（＊は引用文献）

＊ Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16,* 486-514.

＊ Educational Testing Service. (2013). The Official Guide to the TOEFL Test (4th Edition).

＊ Haladyna, T.M., & Downing, S.M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in education, 2,* 37-50.

＊ Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in education, 15,* 309-333.

＊ Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

＊ Iimura, H. (2014). Attractiveness of Distractors in Multiple-Choice Listening Tests. *JLTA Journal, 17,* 19-39.

＊ Ikeno, O. (1996). The effects of text-structure-guiding questions on comprehension of texts with varying linguistic difficulties. *JACET Bulletin, 27,* 51-68.

＊ Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15,* 261-278.

＊ Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19,* 193-220.

＊ 国際ビジネスコミュニケーション協会TOEIC運営委員会（編）. (2016).「公式TOEIC Listening & Reading 問題集 1」. 東京：凸版印刷.

＊ 国際ビジネスコミュニケーション協会TOEIC運営委員会（編）. (2017).「公式TOEIC Listening & Reading 問題集 2」. 東京：凸版印刷.

＊ Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

＊ 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』. 東京：三省堂.

＊ 日本英語検定協会. (2017).『英検2級一次試験過去問』. http://www.eiken.or.jp/eiken/exam/grade_2/solutions.html (2017年7月1日閲覧)

＊ 旺文社（編）. (2003).「英検2級過去6回全問題集」. 東京：大日本印刷.

＊ 旺文社（編）. (2017).「英検2級過去6回全問題集」. 東京：大日本印刷.

＊ 清水真紀. (2005).「リーディングテストにおける質問タイプ―パラフレーズ・推論・テーマ質問と処理レベルの観点から―」. *EIKEN BULLETIN, 17,* 48-62.

＊ Shizuka, T. (1999a). Combining response correctness and confidence level rating to produce polychotomous data from MC dichotomous items. *Japan Colleges of Technology Education Journal, 22,* 243-252.

＊ Shizuka, T. (1999b). Combining response correctness and confidence level rating to produce polychotomous data from MC dichotomous items (Ⅱ). *The Kanto-Koshinetsu Association of Teachers of English (KATE) Bulletin, 13,* 73-86.

＊ Shizuka, T. (2003). The effect of clustered objective probability scoring with truncation (T-COPS) on reliability. In H. Yanai, A. Okada, K. Shigematsu, Y.Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 273-280). Tokyo: Springer.

＊ 髙木修一. (2011).「多肢選択式問題は公平か―記憶保持の強さから―」. *EIKEN BULLETIN, 23,* 13-29.

＊ 卯城祐司（編）. (2012).『英語リーディングテストの考え方と作り方』. 東京：研究社.

＊ Ushiro, Y., Morimoto, Y., Hijikata, Y., Nakagawa, C., Watanabe, F ., Kai, A., Shimizu, H.,Koga, T., Ohno, M., Umehara, C., Hamano, Y., Narumi, T., Tao, N., Shimada, S., Kiyama, M., Suzuki, R., Kurisu, Y., & Gomi, N. (2007). What makes distractors plausible in multiple-choice reading tests? *JLTA Journal, 10,* 56-67.

＊ Ushiro, Y., Nakagawa, C., Morimoto, Y., Hijikata, Y., Watanabe, F., & Kai, A. (2008). Effects of question types on item difficulty in two reading test formats: Open-ended and multiple-choice. *Annual Review of English Language Education in Japan (ARELE), 19,* 201-210.

＊ Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York, NY: Academic Press.

資料1　調査2で用いたテキスト例(英検2級2001年度第2回4Aより)

Green Cleaners

One of the biggest problems facing the world today is how to clean up areas that have been polluted by industry or war. Often these areas contain dangerous metals or chemicals that can poison people or wildlife. Until now, the main way to clean up such areas has been to dig up the polluted earth and take it away. One problem with this method is that it is very expensive. Moreover, this simply moves the pollution to a different place rather than truly solving the problem.

Recently, scientists have been studying plants and trees to see if they can provide a better solution to the pollution problem. People have known for a long time that some plants and trees are good at removing dangerous metals and chemicals from the earth. Scientists are now studying exactly how they do this in order to develop new kinds of plants that may be used to clean up polluted land. Scientists have actually discovered that there are two main ways in which plants help to clean up the earth. One way is by directly absorbing metals and chemicals through their roots and storing them in their branches and leaves. The other way is by attracting insects and other small creatures that can consume dangerous materials in the earth.

Scientists have made important progress with these methods. In America, for example, a kind of grass has been used to clean up areas polluted by oil. But such methods also have some problems. One problem is that they take a long time to work. Even though scientists have developed plants that work more quickly, it still takes many years to clean up an area by using them. Another problem is that when animals eat such plants, they can be poisoned or spread toxic material to other animals or people. Plants cannot be used in every case of pollution. But it seems likely that in the future they will play a major role in helping to repair the damage human beings have caused to nature.

資料2　調査2で用いた多肢選択式問題の例

(注) () 内は選択肢のパラフレージング・レベルを表す。(実施時には無記載)

※ NC = Near Copy, MIR = Minimal Revision, MOR = Moderate Revision, SR = Substantial Revision.

(1) The traditional way to clean up polluted areas

解答			自信度
	1	is an inexpensive means to deal with the problem. (MOR)	低　中　高
	2	can poison human beings and wildlife nearby. (NC)	低　中　高
	3	does not function in places with dangerous metals. (MIR)	低　中　高
	4	does not really have a chance at resolving the issue. (正: SR)	低　中　高

(2) Scientists

解答			自信度
	1	have realized that bugs carry harmful substances. (SR)	低　中　高
	2	are researching the relationship between harmful metals and substances. (MOR)	低　中　高
	3	are trying to use plants to solve the problem of earth pollution. (正: NC)	低　中　高
	4	have found that plants can release poisonous metals. (MIR)	低　中　高

(3) What is one function of plants and trees?

解答			自信度
	1	They utilize metals and substances for growing rapidly. (MOR)	低　中　高
	2	They absorb metals and chemicals and store them inside. (正: MIR)	低　中　高
	3	They keep small creatures away from polluted areas. (NC)	低　中　高
	4	They hold waste underground so that it is less harmful. (SR)	低　中　高

(4) What is one problem of using plants to clean up polluted earth?

解答			自信度
	1	Scientists have not been able to make plants work more quickly. (NC)	低　中　高
	2	It takes a long time to plant a region with plants and trees. (MIR)	低　中　高
	3	There is a possibility that animals will feed the harmful plants. (正: MOR)	低　中　高
	4	Forests are seriously harming the environment globally. (SR)	低　中　高

(5) Which of the following statements is true?

解答			自信度
	1	Plants are acting as a rapid solution to the pollution problem. (MIR)	低　中　高
	2	Military conflicts are one of the sources of soil contamination around society. (正: SR)	低　中　高
	3	Researchers have created plants that do not harm the ground. (MOR)	低　中　高
	4	Animals are more effective at removing pollution than plants. (NC)	低　中　高

B 実践部門 ● 報告Ⅱ ● 英語能力向上をめざす教育実践

CAN-DOリストとCAN-DO Check Sheetの導入が高校生英語学習者の自己調整学習と英語力向上に及ぼす効果

研究者：兵庫県／兵庫県立川西緑台高等学校　教諭 津田 敦子（申請時：神戸大学附属中学教育学校 教諭）
《研究助言者：小池 生夫》

概要

本研究では，4年生（高1）のコミュニケーション英語Ⅰの授業において，自律的学習に効果的であると考えられるCAN-DOリストとCAN-DO Check Sheetを導入し，認知・行動・情意面を包括した概念である自己調整学習と英語力の変化を調査した。4年生全員に自己調整学習に関する質問紙調査を3回実施し，実践を行ったクラスの生徒（実験群）とそれ以外の生徒（統制群）の変化を比較した。その結果実験群の生徒の平均値は統制群の生徒よりも高くなったことから，CAN-DOリストとCAN-DO Check Sheetの活用が，自己調整学習者の育成に有効であることが検証された。また本実践で使用したCAN-DO Check SheetやReflection Sheetの記述内容や自己評価を分析したところ，生徒は自分に厳しくなり過小評価しがちで自己効力感を失っていく可能性があることも明らかになった。そこからCAN-DOリストとCAN-DO Check Sheetを導入する際には，ただ導入するだけでなく「できる感」を実感させる配慮や工夫が必要であるという教育的示唆を得た。

1 はじめに

平成25年に「英語教育改革実施計画」が出されたが，週に5～6時間の授業で高校3年間に身につけられる英語力には限界がある。生徒がそこで述べられているようなレベルに到達するためには，教室外でも更には高校卒業後も英語学習を持続していくような自律した学習者を育成していかなければならない。生徒の英語の4技能を向上させつつ自律した学習者を育成していくために，英語教員は授業改善と指導力向上が迫られている。

「CAN-DOリスト」は「英語教育改革実施計画」の中の「小・中・高等学校を通じて目標・取り扱う内容・評価を改善」の1項目として盛り込まれ，「英語を用いて～することができる」という目標設定（CAN-DOリスト）に対応して4技能を評価することが示された。「CAN-DOリスト」は「実施計画」に先立ち公開された『各中・高等学校の外国語教育における「CAN-DOリスト」の形での学習到達目標設定のための手引き』（文部科学省，2013）の中で，その目的を ①指導と評価の改善 ②4技能の総合的な育成 ③自律的な学習者の育成としており，各校でその目的を達成するよう活用することが求められている。

平成28年度「英語教育実施状況調査」における「CAN-DOリスト」の中学及び高校での設定率を見ると，中学校で75.2%，高校で88.1%と次第にその考え方が浸透してきている様子が分かる。しかし，公表率は中学校12.0%，高校28.4%，到達目標達成状況の把握率は中学校34.2%，高校41.6%にとどまり，その活用はまだ十分とは言えない。

2 先行研究

2.1 自己調整学習

「自己調整学習」とは心理学分野で提唱された概念で，ジマーマンの定義によると「学習者がメタ認知，動機づけ，行動において自分自身の学習過程に能動的に関与していること」(Zimmerman, 1989, p 329) である。自己調整学習は他の自律的学習とは違って学習過程がより分析的に研究されており (辰野，1997)，認知・行動・情意をともに含みこんだ概念であることから，この概念に着目することでより実践的な示唆を与えてくれる (中田，2011)。そこで本研究では自律的学習者を自己調整学習の視点から，つまり認知・行動・情意の3観点で分析する。

2.2 CAN-DOリストを活用した実践研究

2011年に文部科学省が「国際共通語としての英語力向上のための5つの提言と具体的施策」の中で，学習到達目標を「CAN-DO リスト」の形で設定することを提言して以来，中・高等学校でのリスト作成が始まり，様々な実践研究報告がなされてきた。

岡崎 (2014) ではCAN-DOリストに基づいた英語授業に関する高校生の意識調査を実施し，授業全体への満足度，英語力の伸長についての満足度，英語学習への意欲，生徒の理解力の伸びを報告している。兵庫県においても2015年に『「兵庫版基本CAN-DO リスト」及びその活用について』が公表され，その中で拠点校の実践例が報告されている。その中ではCAN-DOリスト活用時の課題や反省点，改善点も言及されていた。渡部 (2015) では「CAN-DO リスト」の作成・開発と授業での活用の様子が報告され，生徒が「CAN-DO リスト」によってモチベーションを高めていくことが報告されている。しかし自己調整学習の観点からはまだ研究がなされていない。

長沼 (2008, 2010) は，CAN-DO尺度の運用に当たっては学習者の動機づけや自己効力感にも配慮する必要があり，スキルと情意の両面のバランスをどのように保っていくかが，今後のCAN-DO研究の鍵といえると述べている。そして『小学校英語Can-Do 評価尺度活用マニュアル解説』(長沼，2017) の中でできる感を感じさせるための留意点を解説している。

そこで本研究では，CAN-DOリストを活用した実践を行い，自己調整学習の観点から高校生英語学習者の変化を分析する。

3 研究の目的

本研究では，自律的学習に効果的であると考えられるCAN-DO リストとCAN-DO Check Sheetを授業に導入し，生徒の情意面と学習方略，英語力の変化を調査する。第1に，英語学習においてCAN-DO リストとCheck Sheetを利用することで生徒が「自己調整学習」を身につけていくことを検証し，その過程を可視化する。第2に，CAN-DO リストとCheck Sheetを利用して「自己調整学習」力を身につけることが実際に英語力の向上につながっていくのか調査する。第3に生徒の英語力によってReflection Sheetの自己評価とCAN-DO Check Sheetの自由記述内容に違いがあるか分析する。そしてこの研究を通して自律した英語学習者を育成するための効果的なCAN-DO リストとCheck Sheetの活用法を提案したいと考える。

4 調査

4.1 リサーチクエスチョン

RQ1 CAN-DO リストを提示しCAN-DO Check Sheetを授業に導入することで，生徒の「自己調整学習」はどのように変化するか？

RQ2 Reflection SheetやCAN-DO Check Sheetの生徒自身の自己評価と自由記述内容はどのように変化するか？

RQ3 CAN-DOリストとCAN-DO Check Sheetの導入が，実際に英語力の向上につながるのか？

RQ4 Reflection Sheetの自己評価とCAN-DO Check Sheetの自由記述内容は生徒の英語力と相関や関連性があるのか？

4.2 実践と調査の手順

4.2.1 対象授業

本実践はコミュニケーション英語Ⅰ（週3時間）の授業で，研究者が担当するクラスで行った。コミュニケーション英語Ⅰの授業は習熟度別少人数授業を実施し，2クラスを3展開，1クラスを2展開にし，計8クラスで授業を行っていた。そのため教師は同時に3人立ち，または2人立ちで同じ授業を展開していた。そしてそのうちの3クラスが研究者の担当するクラスであった。

本校のコミュニケーション英語の授業ではCLILの考え方を取り入れ，単元学習を行っている。そのため考査までの2か月間同じテーマを扱う。昨年度は前期中間考査までが「私の夢」，前期期末考査までが「日本文化（特に漫画アニメといったサブカルチャー）」，後期中間までが「日本の技術革新」，後期期末までが「神戸の震災からの復興」をテーマにした。

4.2.2 対象者

本実践の対象者は4年生（高1）で，4年生は5クラス172名である。そのうち研究者が担当する生徒を実験群，その他の生徒を統制群とした。定期考査ごとにクラス替えをしたので，年間を通して同じ生徒を実験群とすることができず，実験群の生徒は入れ替わっている。

前期（4月～9月）は名列順に単純に分けて少人数クラスを編成しただけであり，各クラスの平均学力はほぼ同じであった。しかしクラス内の学力差は大きかった。前期は中間考査でクラス替えをしなかったので，実験群は半年間同じ生徒であった。後期（10月～3月）は定期考査の成績と本人の希望をもとにAdvancedクラスとBasicクラスを編成し，2度クラス替えを行った。

以下はクラス編成を表にまとめたものである。

■表1：コミュニケーション英語Iの授業クラス編成

HRクラス	前期（4月～9月）	後期の前半（10月～12月）	後期の後半（1月～3月）
1･2組	平等クラス3（1クラスを研究者が担当）	Advanced 1，Basic 2（1クラスを研究者が担当）	Advanced 1，Basic 2（1クラスを研究者が担当）
3組	平等クラス2（1クラスを研究者が担当）	Intermediate 1（研究者が担当），Basic 1	Intermediate 1（研究者が担当），Basic 1
4･5組	平等クラス3（1クラスを研究者が担当）	Advanced 1（研究者が担当），Basic 2	Advanced 1（研究者が担当），Basic 1

4.2.3 実践とデータ収集

2017年4月から2018年3月の1年間，以下のように実践を行い，データを収集した。

①新学期コミュニケーション英語の第1時間目に同時展開の合同クラスでオリエンテーションを実施し，4年生全員に対して本校のCAN-DOリスト（Kobe University Secondary School Framework，以下KUSF）を提示し，年度末での到達目標を提示した。次に春休み課題のスピーキングテストを1人ずつ受けている間に，4年生全員に対して第1回目の「自己調整学習」に関する質問紙調査を実施した。

②研究者は，4月のコミュニケーション英語の授業の第2時間目に新単元を開始したとき，

KUSFの文言をその単元と生徒の現状に合わせて書き換え，単元終了時の到達目標を作成した。そして研究者が担当するクラスにのみそれをワークシート冊子の第1ページにつけ，生徒に提示し説明した。

③前期中間考査直前の単元の最後の授業で，単元のReflection Sheetに振り返りを記入させた。Reflection SheetはCAN-DOリストの項目とその他の授業に関する質問項目（4段階評価），記述の振り返りから成る。

④研究者は前期中間考査後6月に新単元を開始したとき，再度KUSFの文言をその単元と生徒の現状に合わせて書き換え，単元終了時の到達目標を作成した。そして4月同様研究者が担当するクラスにのみそれをワークシート冊子の第1ページにつけ，生徒に提示し説明した。

⑤前期中間考査後から期末考査までの期間（6月から9月まで）は毎時間授業終了時にCAN-DO Check Sheetに自己評価を記入させた。CAN-DO Check Sheetには本単元の到達目標が書かれていて，生徒に4段階で評価させた。また自由記述欄を設けており，そこにはその時間にできたと思うことを記入するよう指示した。

⑥夏休み明けのコミュニケーション英語の第1時間目に同時展開の合同クラスで授業をし，夏休み課題のスピーキングテストを実施した。スピーキングテストを1人ずつ受けている間に，4年生全員に対して第2回目の「自己調整学習」に関する質問紙調査を実施した。

⑦前期期末考査直前の単元の最後の授業で，毎時間記入してきたCAN-DO Check Sheetを見ながら単元のReflection Sheetに記入させた。

※前期が終了した段階で，「自己調整学習」に関する質問紙の①4月と②9月の各質問項目の平均，Reflection Sheetの①6月と②9月の各質問項目の平均を確認したところ，随所で下降が確認されたため，実践をいったん中断した。

※下降の原因をいくつか考え，CAN-DO Check Sheetの改善をしたり，授業内のアクティビティやタスクの工夫をしたり，それらの活動とCAN-DO Check Sheetの活用を関連付けるなどの改善を図ったりして，1月から再度実践を再開した。

⑧研究者は1月に新単元を開始したとき，前期同様単元終了時の到達目標を作成した。そして研究者が担当するクラスにのみそれを生徒に提示し説明した。後期から習熟度別クラスで授業展開をしており，後期中間考査後にも2度目のクラス替えをしたので，各授業クラスは新メンバーでスタートした。

⑨1月から後期期末考査までの期間（1月から3月まで）は毎時間授業終了時にCAN-DO Check Sheetに自己評価を記入させた。

⑩後期期末考査直前の単元の最後の授業で，毎時間記入してきたCAN-DO Check Sheetを見ながら単元のReflection Sheetに記入させた。

⑪後期期末考査終了後に，4年生全員に対して第3回目の「自己調整学習」に関する質問紙調査を実施した。

表2は実践とデータ収集の流れをまとめたものである。

4.2.4 質問紙

質問紙は，以前に公立高校生（1119名）を対象に自己調整学習の大規模調査（Tsuda, 2011）を実施したときのものを使用した。事前調査と先行研究をもとに作成し，予備調査や記述統計を経たものである。記述統計の結果，天井効果床効果のある項目はなく，またクロンバックα値は .955であった。質問45項目（詳細は表3）について，6件法のリカート尺度で答える形式である。そのほかに英語学習を振り返って英語に対する感情の変化や学習方法についての回答を求めた。

4.2.5 本校のCAN-DOリスト（KUSF）

本校では「英語によるコミュニケーション能力を活用し，地球上に生きる人間として，自分の足元から世界を見る視点を持ち，同時に多様化が進む世界の動きの中で人々と協力・共生しながら自分の生き方を選択し，地球的視野で考え行動できる生徒」の育成を目指し，Grade1から卒業後までのスキル運用能力の到達目標を定めている。そして第4学年はおおよそGrade4が目安で学年末までに60％がA2.2に到達することを目標としている。この目標は年度当初にGTEC for STUDENTSの点数や英検の級と対応させながら4年生の生徒全員に提示し説明した。Grade4の英語4技能5領域の到達目標は表4の通りである。

■表2: 実践とデータ収集

年月	実践と背景行事	分析データ
2017.4	[4年生全員] オリエンテーション&スピーキングテスト 本校CAN-DOリストを提示し, 年間目標を伝える。	[4年生全員]「自己調整学習」に関する 質問紙調査①(165名)
2017.4	[実験群] 新単元の授業1時間目に 本校CAN-DOリストをもとに作成した単元の目標を示す。	
2017.6	[実験群] 単元の最後の授業でCAN-DOリストの項目を含む単元のReflection Sheetに振り返りを記入する。 (前期中間考査)	[実験群] CAN-DOリストの項目を含む単元の Reflection Sheet①(65名)
2017.6	[実験群] 新単元の授業1時間目に 本校CAN-DOリストをもとに作成した単元の目標を示す。	
2017.6-9	[実験群] 毎授業終了時に CAN-DO Check Sheetに記入する。	[実験群] CAN-DO Check Sheet①
2017.9 夏休明	[4年生全員] スピーキングテスト	[4年生全員]「自己調整学習」に関する 質問紙調査②(165名)
2017.9	[実験群] 単元の最後の授業でCAN-DOリストの項目を含む単元のReflection Sheetに振り返りを記入する。 (前期期末考査)	[実験群] CAN-DOリストの項目を含む単元の Reflection Sheet②(64名)
2017.10	後期開始 *本実践を中断	
2017.12	[実験群] 単元の最後の授業でCAN-DOリストの項目を含む単元のReflection Sheetに振り返りを記入する。 (後期中間考査) (GTEC for STUDENTS)	[実験群] CAN-DOリストの項目を含む単元の Reflection Sheet③(62名) GTEC for STUDENTSのスコア
2018.1	[実験群] 新単元の授業1時間目に 本校CAN-DOリストをもとに作成した単元の目標を示す。	
2018.1-3	[実験群] 毎授業終了時に CAN-DO Check Sheetに記入する。	[実験群] CAN-DO Check Sheet②
2018.3	[実験群] 単元の最後の授業でCAN-DOリストの項目を含む単元のReflection Sheetに振り返りを記入する。 (後期期末考査)	[実験群] CAN-DOリストをもとに作成した単元の Reflection④(62名)
2018.3		[4年生全員]「自己調整学習」に関する 質問紙調査③(160名)

■表3: 質問45項目の詳細

概念		項目数	参考にした先行研究
学習方略	記憶方略	5	Oxford(1990)のSILL, Maeda(2002)
	認知方略	7	
	補償方略	5	
	メタ認知方略	7	
	情意方略	3	
	社会的方略	2	
動機づけ		6	中田(2001), Hiromori(2006), Yashima(2000)
自己効力感		5	Pintrich, P. R., & De Groot, E. V. (1990) のMSLQ, 中田(2001), Mori(2004)
自律		5	中田(2001)

■表4: Grade 4の到達目標(KUSFからの抜粋)

	聞くこと	読むこと	書くこと	話すこと〔発表〕	やり取りすること
内容（話題場面）	学習した内容に関連した社会的諸問題・身近な話題／簡単な説明	学習した内容に関連した社会的諸問題・身近な話題／多読用図書	学習した内容に関連した社会的諸問題・身近な話題・興味のある話題	学習した内容に関連した社会的諸問題・身近な話題・興味のある話題	学習した内容に関連した社会的諸問題・身近な話題・興味のある話題
素材（対象）	話題に関する説明文や, 少し長めの議論等	まとまりのある説明文, やスピーチ原稿／多読本等	スピーチ原稿等	スピーチ, プレゼンテーション, まとまりのある文章の要約等	【発表者】発表後の聞き手からの予測可能な質問 【聞き手】発表者の発表内容
英語（条件）	簡単なことばでゆっくりはっきりとした発音	メインアイデア, あるいは主張が明確に示されている数パラグラフから成る英文(300語程度)／Penguin Readers Level 3程度の読み物	必要に応じて辞書を用いたり, 参考となる情報, 語彙を与えられれば	パラグラフ構成を明確にした原稿を準備し, 簡単なメモを見ながら	【発表者】聞き手からの想定質問をしっかりと準備した上で 【聞き手】メモを取りながら発表を聞いた上で,
論理性	他者の意見を聞く時には, 文と文の論理的な関係を示すことば(ディスコースマーカー)を意識して, 話し手の主張を正確に理解できる。	他者の意見が書かれたものを読む時には, 文と文の論理的な関係を示すことば(ディスコースマーカー)を意識して, 書き手の主張を正確に理解できる。	ある内容について多くの自分の考えを出した後に, 自分の意見とその理由について, 段落構成を意識しながら1段落を書くことができる。	自分の意見や感想を, 具体的な例を付け加え, 筋道を立てて分かりやすく話すことができる。	自分の意見や感想を, 具体的な例を付け加え, 筋道を立てて分かりやすく話すことができる。
行動（理解）	情報を整理しながら聞き分け, 概要を理解することができる。	・辞書を使いながら読めば, 話の主題や要点をほぼ理解することができる。 ・パラグラフ構成を意識し, 主題文と支持文を区別することができる。	基本的な語彙, 構文, よく使われる結束表現(接続詞, 副詞等)を使って, つながりのある文章を書くことができる。	ペア・グループまたは個人で5～10分程度, 比較的自然な言葉遣いで, 伝えたいことをある程度的確に発表することができる。	【発表者】単純な英文で明確に答え, 追加情報を述べることができる。 【聞き手】発表内容を英文でまとめることができる。簡単な質問をすることができる。
限界域	繰り返し発話をしてもらわなければ理解できないこともある。	授業で扱っていない話題は理解に限りがある。 ある文章の一節を理解するために, 読み返したり, 文法書を参照したりする必要がある場合もある。	あまり使われない構文や表現などは, 書き誤ったり, ぎこちない表現になったりすることもある。	・話し方は, 準備されている内容については比較的流暢であるが, 即座の応答に対しては言葉に詰まることが頻繁にある。 ・基本的な誤り(動詞の時制など)が多く起こり, 時に理解されないこともある。	・話し方は, 準備されている内容については比較的流暢であるが, 即座の応答に対しては言葉に詰まることが頻繁にある。 ・基本的な誤り(動詞の時制など)が多く起こり, 時に理解されないこともある。

4.2.6 本実践で使用した単元CAN-DOリスト

KUSFの内容は細かく，教師側のみが意識すればよい内容も含まれているため，単元CAN-DOリストはそこから生徒が意識すべき部分のみを抜粋し，生徒にわかり易くシンプルなものを心掛けた。以下は最終版で後期後半1月~3月の単元で生徒に提示したものである。話すことはKUSFで「ペア・グループまたは個人で5~10分」となっているが，単元の最後に4人グループで10分程度のプレゼンテーションをさせることとし，個人では2分程度発表できればよいと設定した。

4.2.7 本実践で使用したCAN-DO Check Sheet

図1の単元CAN-DOリストに対応させて，Can-Do Check Sheetを作成し，毎授業最後の2分程度で生徒に記入をさせた。図2が前期後半6月～9月まで使用したものである。生徒は各項目を4段階で評価したのち，下欄（実際のものはかなり幅を取っている）に上段項目以外にその授業内に「できた」と思うことを書くように伝えた。

UNIT 4　Recovery from the Great Hanshin Earthquake 「神戸の復興」から考える— Resilient Community を目指して		
Research Question		1. Has Kobe become an adequately disaster resilient community? 2. What can Kobe teach other cities about becoming disaster resilient?
レポート＆プレゼンテーションの内容		1．提示されたテーマと英文の中からグループで調べる項目を1つ選定する。 2．英文内容とリサーチ結果をまとめる。 3．Research Question に対する考察をまとめる。 4．まとめた内容を，発表する。（グループで 10 分程度） ＊クラス内発表→優秀グループ選出→学年発表
目標	聞くこと	情報を整理しながら聞き分け，話し手の主張や概要を理解することができる。
	読むこと	辞書を使いながら読めば，話の主題や要点をほぼ理解することができる。
	書くこと	①自分の意見とその理由を，構成を意識しながら1段落を書くことができる。 ②基本的な語彙，構文，よく使われる結束表現（接続詞，副詞等）を使って，つながりのある文章を書くことができる。"
	話すこと	①本文の内容を，筋道を立てて分かりやすく話すことができる。（まとまりのある文章の要約） ②パラグラフ構成を明確にした原稿を準備し，簡単なメモを見ながら，自分の意見や感想を，具体的な例を付け加え，筋道を立てて分かりやすく話すことができる。（スピーチ，プレゼンテーション等） ③個人で2分程度，比較的自然な言葉遣いで，伝えたいことをある程度的確に発表することができる。
	やり取り	【発表者】想定質問の準備をした上で，聞き手からの質問に対して単純な英文で明確に答え，追加情報を述べることができる。 【聞き手】①発表内容を英文でまとめることができる。 ②メモを取りながら発表を聞いた上で，簡単な質問をすることができる。

■図1：本実践の後期後半1月～3月に使用した単元CAN-DOリスト

前期が終了した時点で質問紙やReflection Sheetのデータを集計したところ，多くの項目で数値が下がっていたことから，Can-Do Check Sheetも改訂を加えた。図3が後期後半1月～3月まで使用したものである。上段はその授業内にできたと思う項目について4段階評価をして1～4までの数字を書きこむ。下欄（実際のものはかなり幅を取っている）は生徒のJournalとして様々な使い方をした。

点線の上は，授業開始時に毎時間ペアでSmall Talkをするので，その直後に「新たに知った単語やフレーズ」「言いたいけど言えなかった表現」

UNIT 2 Cool Japan **Can-Do Check Sheet**

Class **No.** **Name**

4:（75％以上） 3:（50%～75％） 2:（25%～50％） 1:（25％以下）

Can-Do リスト		回	1	2	3	4	5	6	7	8	9	10
		DATE	/	/	/	/	/	/	/	/	/	/
聞くこと	教師による Oral Introduction または Summary を聞いて，情報を整理しながら聞き分け，概要を理解することができる。		4	4	4	4	4	4	4	4	4	4
			3	3	3	3	3	3	3	3	3	3
			2	2	2	2	2	2	2	2	2	2
			1	1	1	1	1	1	1	1	1	1
読むこと	辞書を使いながら読めば，話の主題や要点をほぼ理解することができる。		4	4	4	4	4	4	4	4	4	4
			3	3	3	3	3	3	3	3	3	3
			2	2	2	2	2	2	2	2	2	2
			1	1	1	1	1	1	1	1	1	1
書くこと	①自分の意見とその理由を，１段落を書くことができる。 ②つながりのある文章を書くことができる。		4	4	4	4	4	4	4	4	4	4
			3	3	3	3	3	3	3	3	3	3
			2	2	2	2	2	2	2	2	2	2
			1	1	1	1	1	1	1	1	1	1
話すこと	①本文の内容や自分の意見や感想を，筋道を立てて分かりやすく話すことができる。 ②個人で（2 分程度）伝えたいことをある程度的確に発表することができる。		4	4	4	4	4	4	4	4	4	4
			3	3	3	3	3	3	3	3	3	3
			2	2	2	2	2	2	2	2	2	2
			1	1	1	1	1	1	1	1	1	1
やり取り	【発表者】聞き手からの質問に対して単純な英文で明確に答え，追加情報を述べることができる。 【聞き手】簡単な質問をすることができる。		4	4	4	4	4	4	4	4	4	4
			3	3	3	3	3	3	3	3	3	3
			2	2	2	2	2	2	2	2	2	2
			1	1	1	1	1	1	1	1	1	1

回	Date	What more can you do?
	/	
	/	
（中略）		
	/	

■図2：本実践の前期後半6月～9月に使用したCAN-DO Check Sheet

を記入させた。点線の下は，Basic クラスでは日本語で「感想」「わからなかった点」「教師に伝えたいこと」など自由に書かせて教師とのやり取りに使用した。Advanced クラスでは授業内で扱った内容について感想や意見などを英語で数行書かせたりした。心掛けたことは生徒の自由度を増やす、生徒のレベルに応じて利用の仕方を変える、生徒にできた感を持たせる、記入することのメリットを実感させることであった。

UNIT 4 Recovery from the Great Hanshin Earthquake Can-Do Check Sheet

Class No. Name

4:（75％以上） 3:（50%～75％） 2:（25%～50％） 1:（25％以下）

Can-Do リスト			回	1	2	3	4	5
			DATE	/	/	/	/	/
聞くこと	情報を整理しながら聞き分け，話し手の主張や概要を理解することができた。							
読むこと	話の主題や要点をほぼ理解することができた。							
書くこと	自分の意見とその理由について，つながりのある文章を書くことができた。"							
話すこと	本文の内容や，自分の意見や感想を，筋道を立てて分かりやすく話すことができた。							
やり取り	【発表者】	聞き手からの質問に対して単純な英文で明確に答え，追加情報を述べることができた。						
	【聞き手】	簡単な質問をすることができた。						

回	Date	Journal
	/	
	/	
	/	
（中略）		
	/	

■図3：本実践の後期後半1月～3月に使用したCAN-DO Check Sheet

4.2.8 本実践で使用したReflection Sheet

考査直前の各単元の最終授業で以下のReflection Sheet（図4）に記入させて，その単元の振り返りをさせた。Reflection Sheet 内の1には図1の内容を再度提示し，Reflection Sheet 内の2にはその期間に実施した主なチェックやテストの教師がつけた評価を，確認の意味も含めて記入して，生徒に渡した。Reflection Sheet 内の3は質問項目が25問あり，1～7が単元CAN-DOリストの項目で，その他の授業に関する質問項目が続く。Reflection Sheet 内の4は振り返りを自由記述させた。

UNIT 2　CoolJapan　　Reflection Sheet

1. 単元目標（省略）

2. Evaluation

2-1 Miyazaki	2-2 Tezuka	2-3 What makes manga～?		単語テスト	
Summary	Summary	Summary	Presentation	①	②

3. 自己評価

4:（75%以上）　3:（50%～75%）　2:（25%～50%）　1:（25%以下）

1.【聞くこと】 教師による Oral Introduction または Summary を聞いて、情報を整理しながら聞き分け、概要を理解することができる。 (4 – 3 – 2 – 1)

2.【読むこと】辞書を使いながら読めば、話の主題や要点をほぼ理解することができる。 (4 – 3 – 2 – 1)

3.【書くこと】本文の内容やマンガ・アニメについて、自分の意見とその理由を、段落構成を意識しながら１段落を書くことができる。 (4 – 3 – 2 – 1)

4.【書くこと】基本的な語彙、構文、よく使われる結束表現（接続詞、副詞等）を使って、つながりのある文章を書くことができる。 (4 – 3 – 2 – 1)

5.【話すこと】　絵を見ながら、本文の内容を筋道を立てて分かりやすく話すことができる。（Summary） (4 – 3 – 2 – 1)

6.【話すこと】パラグラフ構成を明確にした原稿を準備し、簡単なメモを見ながら、自分の意見や感想を、具体的な例を付け加え、筋道を立てて分かりやすく話すことができる。（プレゼンテーション） (4 – 3 – 2 – 1)

7.【話すこと】個人で 2 分程度、比較的自然な言葉遣いで、伝えたいことをある程度的確に発表することができる。 (4 – 3 – 2 – 1)

（以下省略）

4. 振り返り

■図4：Reflection Sheet

5 結果と考察

5.1 (RQ1)CAN-DOリストを提示しCAN-DO Check Sheetを授業に導入することで,生徒の「自己調整学習」はどのように変化するか？

5.1.1 分析①（前期4月～9月）

前期が終了した時点で「自己調整学習」質問紙調査①（4月）と②（9月）のデータを集計したものが表５である。

4月から9月にかけて多くの項目で数値が下がっており，一見すると統制群よりも実験群のほうが下降が著しいとも取れる結果であった。CAN-DOリストを生徒と共有し，生徒に英語学習の明

■表5:「自己調整学習」質問紙調査①(4月)と②(9月)の結果

No.	質問項目	実験群(n=61)				統制群(n=104)			
		4月	9月	差	増加率	4月	9月	差	増加率
＊Memory Strategies									
1	英語ですでに知っていることと新しく学習していることとの関係を考える	3.8	3.9	0.06	0.02	4.0	4.0	0.04	0.01
2	覚えやすいように文の中で新しい語を覚える	4.1	4.0	- 0.15	- 0.04	4.3	3.9	- 0.36	- 0.08
3	新しい単語を覚えるために, 発音と単語が指すもののイメージや外見を結び付けて考える	4.1	4.1	- 0.01	- 0.00	4.2	4.1	- 0.08	- 0.02
4	単語が使われる場面を心に想像して新しい単語を覚える	3.8	3.6	- 0.14	- 0.04	3.8	3.7	- 0.12	- 0.03
5	英語の授業の復習をよくする	3.4	3.3	- 0.11	- 0.03	3.4	3.3	- 0.04	- 0.01
＊Cognitive Strategies									
6	新しい単語を数回書いたり言ったりする	4.2	3.8	- 0.42	- 0.10	4.3	4.1	- 0.19	- 0.04
7	英語のネイティブスピーカーのように話そうとする	4.0	3.7	- 0.30	- 0.08	4.2	4.1	- 0.11	- 0.03
8	英語の発音練習をする	3.7	3.7	- 0.00	- 0.00	4.0	3.8	- 0.17	- 0.04
9	英語の文章にざっと目を通し, その後前に戻って注意深く読む	4.1	4.0	-0.15	- 0.04	4.0	4.0	0.03	0.01
10	英語の中にパターンを見つけようとする	3.6	3.7	0.13	0.04	3.8	3.7	- 0.10	- 0.03
11	英単語を分解して, 意味を知ろうとする	3.9	3.6	- 0.35	- 0.09	3.8	3.8	0.03	0.01
12	英語の1語1語を日本語に直して, 英文を読む	3.3	3.3	0.02	0.01	3.5	3.6	0.11	0.03
＊Compensation Strategies									
13	知らない単語を理解しようと推測する	4.4	4.3	- 0.16	- 0.04	4.3	4.3	- 0.09	- 0.02
14	英語で会話をしていて適切な語が思いつかない時, ジェスチャーを使う	4.5	4.3	- 0.19	- 0.04	4.4	4.6	0.16	0.04
15	英語で適切な語がわからない時, 知っている単語を組み合わせて表現する	4.7	4.5	- 0.21	- 0.04	4.4	4.5	0.15	0.03
16	英語を読む時, 文の前後の意味や状況から大まかな意味を考えようとする	4.8	4.5	- 0.22	- 0.05	4.5	4.4	- 0.04	- 0.01
17	英語の単語が思いつかない時, 似た意味の別の単語を使おうとする	4.6	4.4	- 0.22	- 0.05	4.5	4.6	0.07	0.02
＊Metacognitive Strategies									
18	自分の英語の間違いに気づき, そこから学んで上達しようとする	4.2	4.1	- 0.08	- 0.02	4.1	4.1	- 0.02	- 0.00
19	他の人が英語を話している時, 集中して聞く	4.4	4.3	- 0.16	- 0.04	4.3	4.2	- 0.17	- 0.04
20	優れた英語学習者になるためにどうしたらよいか考える	4.0	3.7	- 0.30	- 0.08	3.7	3.6	- 0.11	- 0.03
21	英語の学習計画を立てる	3.5	3.4	- 0.04	- 0.01	3.5	3.4	- 0.09	- 0.03
22	英語の学習に十分時間をあてる	3.7	3.7	0.04	0.01	3.8	3.6	- 0.18	- 0.05
23	英語の能力を高めるための目標を持っている	4.0	3.6	- 0.34	- 0.09	4.2	3.8	- 0.43	- 0.10
24	自分の英語学習を振り返り, うまく学習が進められているかどうか考える	3.5	3.4	- 0.17	- 0.05	3.6	3.5	- 0.18	- 0.05

■表5:「自己調整学習」質問紙調査①(4月)と②(9月)の結果

No.	質問項目	実験群(n=61)				統制群(n=104)			
		4月	9月	差	増加率	4月	9月	差	増加率
＊Affective Strategies									
25	間違いを恐れず英語を話すよう自分を励ます	3.8	3.5	- 0.27	- 0.07	3.7	3.4	- 0.27	- 0.07
26	うまくいった時, 自分をほめる	3.7	3.5	- 0.24	- 0.06	3.7	3.6	- 0.08	- 0.02
27	英語を使っている時は, 緊張している	4.4	4.2	- 0.21	- 0.05	4.4	4.1	- 0.24	- 0.06
＊Social Strategies									
28	英語がわからない時, ゆっくり話してもらうか, もう１度言ってもらう	4.6	4.5	- 0.12	- 0.03	4.6	4.6	0.06	0.01
29	英語を話す人の文化を学ぶよう心掛ける	3.8	3.7	- 0.17	- 0.04	3.8	3.7	- 0.15	- 0.04
＊Intrinsic Motivation									
30	私が英語を勉強するのは英語が好きだからだ	3.7	3.6	- 0.14	- 0.04	3.7	3.3	- 0.37	- 0.10
31	英語を勉強していると, とても満足感がある	3.8	3.3	- 0.47	- 0.12	3.8	3.5	- 0.37	- 0.10
32	英語は一生勉強し続けていきたい	4.1	3.9	- 0.20	- 0.05	4.1	3.9	- 0.28	- 0.07
33	高校を卒業すれば(あるいは入試が終われば) 英語は勉強しない(逆転項目)	2.6	2.4	- 0.19	- 0.07	2.4	2.6	0.14	0.06
34	英語を勉強する目的は特にない(逆転項目)	2.8	2.5	- 0.34	- 0.12	2.3	2.6	0.31	0.14
35	英語を勉強する意義が見出せない(逆転項目)	2.4	2.3	- 0.19	- 0.08	2.2	2.5	0.25	0.12
＊Self-efficacy									
36	授業で出された問題や課題を自分はうまくこなせる	3.9	3.7	- 0.22	- 0.06	3.6	3.5	- 0.10	- 0.03
37	教えられる内容を, 自分は理解できる	4.0	3.8	- 0.18	- 0.05	4.1	3.7	- 0.38	- 0.09
38	私は努力すれば, 英語ができるようになる	4.2	4.0	- 0.25	- 0.06	4.2	3.9	- 0.35	- 0.08
39	私は日本語が話せるのだから, 英語もできるようになる	4.1	3.6	- 0.50	- 0.12	3.9	3.4	- 0.51	- 0.13
40	誰にも外国語を習得する能力がある	4.3	4.2	- 0.08	- 0.02	4.2	4.0	- 0.24	- 0.06
＊Autonomy									
41	英語学習において目標を作ることができる	3.8	3.7	- 0.10	- 0.03	3.9	3.7	- 0.23	- 0.06
42	自分なりの英語の学習方法を持っている	3.8	3.5	- 0.32	- 0.09	3.7	3.5	- 0.28	- 0.08
43	明確な英語学習の目的がある	3.7	3.8	0.00	0.00	4.0	3.6	- 0.39	- 0.10
44	自分のペースで英語を学習できる	4.0	3.7	- 0.37	- 0.09	3.9	3.6	- 0.32	- 0.08
45	どう勉強したら成績があがるのかわかっている	3.3	3.1	- 0.20	- 0.06	3.4	3.2	- 0.20	- 0.06

確な目標を持ってもらい，それを生徒自身が振り返ることで，生徒は「できる感」を高め，自己調整学習力や自己効力感を獲得していくであろうという予想のもとに行った実践であったので，この結果にはショックを受け，見直しを迫られた。

一般的に高校レベルになって学習内容が難しくなり，英語が嫌いになったり，自己効力感が低下したりしていく傾向があることから推測すると，やむ得ない結果なのかもしれない。また様々な意識調査において，一般的に本校CAN-DOリストのような高い目標を示すと，自分に厳しくなり低い評価を下すようになる傾向も指摘されている。

「38 私は努力すれば，英語ができるようになる」「39 私は日本語が話せるのだから，英語もできるようになる」などが下がっている様子を見ると，この半年で自己効力感が低下したことがうかがえた。ただこれらの項目に関しては実験群も統制群も同様に低下していることから，CAN-DOリスト（KUSF）やCAN-DO Check Sheetが原因というよりは，高校レベルになって学習内容が難しくなったために，自己効力感を低下させていったのではないかと考えられた。「26 うまくいった時，自分をほめる」「31 英語を勉強していると，とても満足感がある」「36 授業で出された問題や課題を自分はうまくこなせる」等が，研究者のクラスで大きく下がっているのは，CAN-DOリスト（KUSF）を提示し，CAN-DO Check Sheetで振り返りをさせることで，生徒は自分に厳しくなり結果として自己効力感を失う結果となったことが予測された。また「20 優れた英語学習者になるためにどうしたらよいか考える」「24 自分の英語学習を振り返り，うまく学習が進められているかどうか考える」「44 自分のペースで英語を学習できる」などは毎時間Can-Do Check Sheetで振り返りを行ったにもかかわらず，自分の学習に対するメタ認知力も学習者自律も育っていないようであった。これはCAN-DO Check Sheetが生徒にとってやらされているだけのものであって，実際には自分たちへの意義や効果を見出していない結果であろうと考えられた。

次に各項目の平均の差が統計的に有意か確かめるために，有意水準5%で両側検定のt検定を行った。その結果が表6の通りである。

統計的に分析すると下降に有意差が確認されたのは，実験群では1項目のみで，統制群では8項目に及んだ。特に「明確な英語学習の目的がある」という項目に関しては，統制群は有意に下降したのに対して，実験群は全く下降していない。統制群と比較してみると，CAN-DOリストで目標を明確にし，CAN-DO Check Sheetで振り返りを行ったことが何らかの効果を発揮していたのかもしれないと考えられた。ただこの時点では残念ながら期待していたような有意に上昇した項目は皆無であった。その原因が，高校レベルになって学習内容が難しくなり自己効力感を失っていったことに

■表6:「自己調整学習」質問紙調査①（4月）と②（9月）で有意差が確認された項目

	No.	質問項目	4月	9月	差
実験群	39	私は日本語が話せるのだから，英語もできるようになる	4.1	3.6	- 0.50*
統制群	2	覚えやすいように文の中で新しい語を覚える	4.3	3.9	- 0.36*
	23	英語の能力を高めるための目標を持っている	4.2	3.8	- 0.43*
	30	私が英語を勉強するのは英語が好きだからだ	3.7	3.3	- 0.37*
	31	英語を勉強していると，とても満足感がある	3.8	3.5	- 0.37*
	37	教えられる内容を，自分は理解できる	4.1	3.7	- 0.38*
	38	私は努力すれば，英語ができるようになる	3.9	3.4	- 0.51*
	39	私は日本語が話せるのだから，英語もできるようになる	3.9	3.4	- 0.51*
	43	明確な英語学習の目的がある	4.0	3.6	- 0.39*

* $p < .05$, ** $p < .01$, *** $p < .001$

あるのか，高い目標を提示されたために自分に厳しくなり結果として自己効力感を失っていったことにあるのかはわからないが，「自己調整学習」にかかわる項目が全く有意に上昇していなかったので，何らかの改善をしなければならなかった。

5.1.2 分析②（後期9月～3月）

いったんこの実践を中断し再考したのち，単元のCAN-DOリストの提示の仕方，CAN-DO Check Sheetの内容や利用の仕方を改善し（4.2.7参照），後期の後半（1月～3月）に実践を再開した。また授業内のアクティビティやタスクがCAN-DO Check Sheetとできるだけ連動するよう意識した。

後期が終了した時点で「自己調整学習」質問紙調査②（9月）と③（3月）のデータを分析した。10月と1月にクラス替えをしたため，分析②における実験群は後期後半（1月～3月）に研究者の担当クラスに所属していた生徒であり，分析①の実験群の生徒とはメンバーが入れ替わっている。また実践は1月～3月であるが，全員に対する質問紙調査を行ったのが9月と3月であるため，期間が完全には一致していない。実験群では45項目中34項目で平均の上昇が見られた。しかし有意水準5%で両側検定のt検定を行ったところ，有意に上昇した項目はなかった。一方統制群は33項目が下降しており，「5 英語の授業の復習をよくする」（$p < .05$），「22 英語の学習に十分時間をあてる」（$p < .01$）において有意に下降していた。

質問紙調査③（3月）の実験群（n=59）と統制群（n=101）の各項目の平均について有意水準5%で両側検定のt検定を行った結果，表7の項目において有意差が確認された。

実験群の各項目の平均の上昇に有意差は確認できなかったものの，統制群との比較において有意差が確認されたことから，CAN-DOリストとCAN-DO Check Sheetの導入が自己調整学習の向上に効果があったと結論付けられる。実験群と統制群では，授業者も異なっている上に，後期から習熟度別クラス編成にしたため生徒の質も若干異なっていた可能性はある。また研究者がこの実践をしつつ授業改善を試みていたため，CAN-DOリストとCAN-DO Check Sheetの導入以外の要因もあったと考えられる。しかしCAN-DOリストとCAN-DO Check Sheetの導入を通して授業者が授業改善をすることができていれば，それも間接的に「自己調整学習者」の育成に効果的であったと言えるだろう。

■表7:「自己調整学習」質問紙調査③（3月）の実験群と統制群で有意差が確認された項目

No.	質問項目	実験群	統制群	有意差
ハ	（全般的な質問）英語の勉強は学び続けていきたい。	4.3	3.8	**
10	英語の中にパターンを見つけようとする	4.1	3.5	**
14	英語で会話をしていて適切な語が思いつかない時，ジェスチャーを使う	4.8	4.3	**
16	英語を読む時，文の前後の意味や状況から大まかな意味を考えようとする	4.8	4.4	*
17	英語の単語が思いつかない時，似た意味の別の単語を使おうとする	4.8	4.4	*
25	間違いを恐れず英語を話すよう自分を励ます	3.7	3.2	*
26	うまくいった時，自分をほめる	3.8	3.4	*
43	明確な英語学習の目的がある	3.9	3.4	**

* $p < .05$, ** $p < .01$, *** $p < .001$

5.1.3 分析③（年間4月～3月）

次にCAN-DOリストとCAN-DO Check Sheetを活用した生徒が自己調整学習の観点から年間でどのように変化したかを調べるため，CAN-DOリストとCAN-DO Check Sheetの実践を行った期間（6月～9月と1月～3月の両方）に研究者のクラスに所属していた生徒を実験群として抽出し分析した。比較をするため，上記期間

のどちらにも研究者のクラスに所属していなかった生徒を統制群とした。

以下（表8）がその結果である。

■表8:「自己調整学習」質問紙調査①（4月）②（9月）③（3月）の実験群と統制群の変化

No.	質問項目	実験群（n=31）			統制群（n=77）		
		4月	9月	3月	4月	9月	3月
＊全般的な質問							
イ	英語を学ぶことは楽しい。	4.2		4.0	4.2		3.6
ロ	英語を学ぶことは楽ではないが，満足感がある。	4.3		4.4	4.3		3.8
ハ	英語の勉強は学び続けていきたい。	4.9		5.0	4.7		4.5
＊Memory Strategies							
1	英語ですでに知っていることと新しく学習していることとの関係を考える	3.8	3.9	3.9	4.0	4.0	3.6
2	覚えやすいように文の中で新しい語を覚える	4.1	4.0	4.3	4.1	3.9	3.8
3	新しい単語を覚えるために，発音と単語が指すもののイメージや外見を結び付けて考える	4.2	4.1	4.5	4.1	4.0	4.1
4	単語が使われる場面を心に想像して新しい単語を覚える	3.7	3.7	4.1	3.7	3.5	3.6
5	英語の授業の復習をよくする	3.0	3.1	3.4	3.4	3.3	2.8
＊Cognitive Strategies							
6	新しい単語を数回書いたり言ったりする	4.2	3.6	4.1	4.3	4.1	4.0
7	英語のネイティブスピーカーのように話そうとする	3.9	3.7	4.0	4.1	4.1	3.9
8	英語の発音練習をする	3.6	3.5	3.9	4.0	3.8	3.8
9	英語の文章にざっと目を通し，その後前に戻って注意深く読む	4.0	4.0	4.4	4.0	4.2	3.8
10	英語の中にパターンを見つけようとする	3.5	3.6	4.2	3.7	3.5	3.4
11	英単語を分解して，意味を知ろうとする	4.0	3.5	3.9	3.6	3.7	3.8
12	英語の1語1語を日本語に直して，英文を読む	3.2	3.3	3.5	3.5	3.6	3.4
＊Compensation Strategies							
13	知らない単語を理解しようと推測する	4.4	4.3	4.6	4.3	4.2	4.2
14	英語で会話をしていて適切な語が思いつかない時，ジェスチャーを使う	4.8	4.3	5.0	4.3	4.5	4.4
15	英語で適切な語がわからない時，知っている単語を組み合わせて表現する	4.9	4.5	5.1	4.2	4.4	4.6
16	英語を読む時，文の前後の意味や状況から大まかな意味を考えようとする	4.9	4.5	4.9	4.4	4.4	4.4
17	英語の単語が思いつかない時，似た意味の別の単語を使おうとする	4.9	4.3	4.9	4.3	4.5	4.4
＊Metacognitive Strategies							
18	自分の英語の間違いに気づき，そこから学んで上達しようとする	4.1	4.2	4.5	4.0	4.1	4.0
19	他の人が英語を話している時，集中して聞く	4.6	4.4	4.4	4.2	4.2	4.0
20	優れた英語学習者になるためにどうしたらよいか考える	4.0	3.6	3.7	3.7	3.5	3.5
21	英語の学習計画を立てる						

■表8:「自己調整学習」質問紙調査①(4月)②(9月)③(3月)の実験群と統制群の変化

No.	質問項目	実験群(n=31)			統制群(n=77)		
		4月	9月	3月	4月	9月	3月
22	英語の学習に十分時間をあてる	3.8	3.6	3.8	3.8	3.7	3.2
23	英語の能力を高めるための目標を持っている	4.1	3.5	3.8	4.1	3.7	3.5
24	自分の英語学習を振り返り, うまく学習が進められているかどうか考える	3.5	3.4	3.7	3.6	3.5	3.2
＊Affective Strategies							
25	間違いを恐れず英語を話すよう自分を励ます	3.8	3.5	3.8	3.5	3.3	3.1
26	うまくいった時, 自分をほめる	3.9	3.4	3.7	3.6	3.5	3.3
27	英語を使っている時は, 緊張している	4.5	4.3	4.5	4.4	4.1	4.1
＊Social Strategies							
28	英語がわからない時, ゆっくり話してもらうか, もう1度言ってもらう	4.8	4.6	5.0	4.6	4.6	4.6
29	英語を話す人の文化を学ぶよう心掛ける	3.7	3.9	4.0	3.7	3.6	3.5
＊Intrinsic Motivation							
30	私が英語を勉強するのは英語が好きだからだ	3.7	3.5	3.3	3.6	3.3	3.0
31	英語を勉強していると, とても満足感がある	3.9	3.4	3.8	3.8	3.4	3.2
32	英語は一生勉強し続けていきたい	4.2	3.9	4.2	4.0	3.8	3.6
33	高校を卒業すれば(あるいは入試が終われば)英語は勉強しない(逆転項目)	2.3	2.5	2.7	2.5	2.7	2.7
34	英語を勉強する目的は特にない (逆転項目)	2.6	2.5	2.4	2.4	2.7	2.6
35	英語を勉強する意義が見出せない (逆転項目)	2.2	2.2	2.3	2.2	2.5	2.5
＊Self-efficacy							
36	授業で出された問題や課題を自分はうまくこなせる	3.9	3.6	3.4	3.5	3.5	3.4
37	教えられる内容を, 自分は理解できる	4.0	4.0	4.2	4.0	3.7	3.8
38	私は努力すれば, 英語ができるようになる	4.3	4.0	4.2	4.2	3.9	3.9
39	私は日本語が話せるのだから, 英語もできるようになる	4.1	3.6	3.8	3.9	3.5	3.4
40	誰にも外国語を習得する能力がある	4.2	4.2	4.0	4..3	4.1	3.8
＊Autonomy							
41	英語学習において目標を作ることができる	4.0	3.8	3.8	3.8	3.7	3.5
42	自分なりの英語の学習方法を持っている	3.7	3.4	3.5	3.6	3.5	3.2
43	明確な英語学習の目的がある	3.9	3.8	4.0	3.8	3.5	3.3
44	自分のペースで英語を学習できる	4.0	3.6	3.8	3.8	3.6	3.5
45	どう勉強したら成績があがるのかわかっている	3.2	2.8	3.0	3.3	3.1	3.1

以下は，上記表の分野別の平均をグラフにしたものである。図5は項目1〜29の学習ストラテジーに関する項目をまとめ，図6はそこからメタ認知ストラテジーに関する項目（18〜24）を抽出したものである。図7はIntrinsic Motivation（30〜35＊逆転項目は数値を変換），図8はSelf-efficacy（36〜40），図9はAutonomy（41〜45）をまとめた。

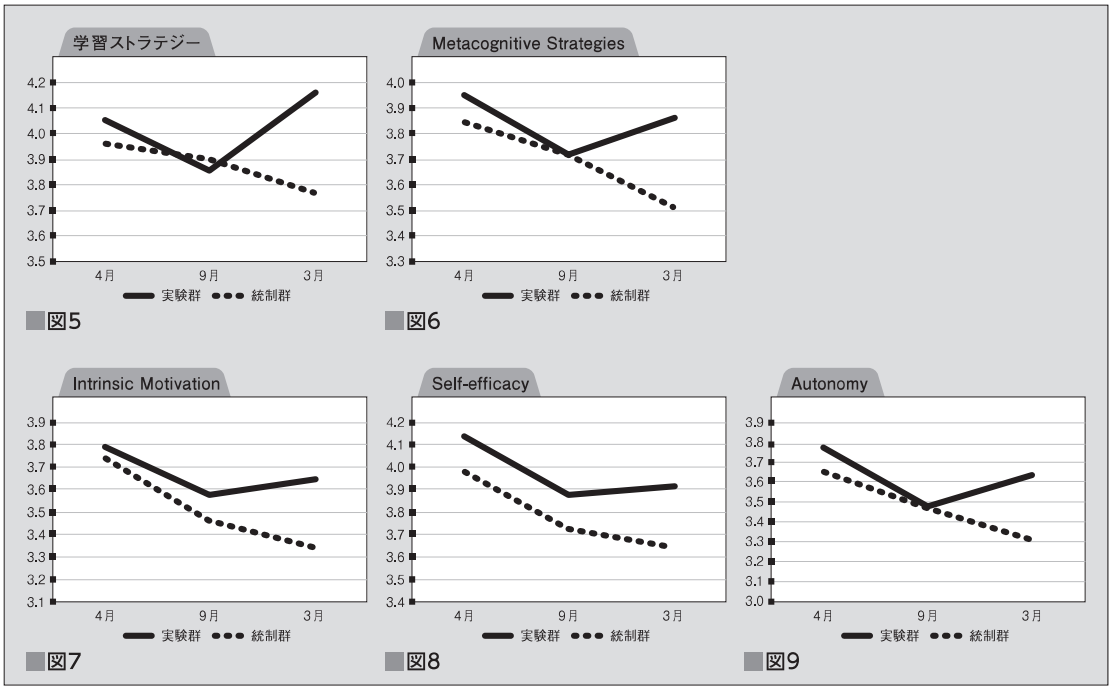


図5

図6

図7

図8

図9

実験群では4月から9月にかけて下降した項目が多いが，ほとんどの項目において9月から3月にかけて上昇していた。有意水準5%で両側検定のt検定を行った結果，4月〜9月に「17 英語の単語が思いつかない時，似た意味の別の単語を使おうとする」が0.05水準で下降していたものの，9月〜3月に「14 英語で会話をしていて適切な語が思いつかない時，ジェスチャーを使う（$p < .01$）」「15 英語で適切な語がわからない時，知っている単語を組み合わせて表現する（$p < .01$）」「17 英語の単語が思いつかない時，似た意味の別の単語を使おうとする（$p < .01$）」，4月〜3月では「10 英語の中にパターンを見つけようとする（$p < .05$）」が有意に上昇していた。 一方統制群では，有意に上昇した項目はなく，ほとんどの項目において下降したままであった。

年間を通して見ると，CAN-DOリストとCAN-DO Check Sheetの導入は，統制群の生徒のように自己調整学習を低下させなかったという点では有効であったと言えるが，4月当初と比べて上昇したという結果を得ることはできなかった。今後はなぜ前半に下降したのか，どうすれば下降を防ぐことができるのかをさらに研究をしていかなければならない。

5.2 (RQ2) Reflection SheetやCAN-DO Check Sheetの生徒自身の自己評価と自由記述内容はどのように変化するか？

前期はクラス替えをしなかったので，半年間担当生徒は同じであったことから，6月と9月のReflection Sheetの分析を行った。以下（表9）に単元の目標（CAN-DOリスト）として提示した項目（1〜7）とその他の質問項目で有意差の出たもの（11,14）を示す。

単元CAN-DOリストでは1と6が有意に上昇していた。2〜4の【読むこと】【書くこと】に関しては6月の数値から高かったので，これ以上の上昇

■表9: 6月と9月のReflection Sheetの分析結果

No.	質問項目	6月	9月	差	増加率	有意差
1	【聞くこと】教師によるOral Introduction または Summary を聞いて，情報を整理しながら聞き分け，概要を理解することができる。	2.89	3.14	0.25	0.09	**
2	【読むこと】辞書を使いながら読めば，話の主題や要点をほぼ理解することができる。	3.47	3.32	- 0.15	-0.04	
3	【書くこと】本文の内容について，自分の意見とその理由を，段落構成を意識しながら1段落を書くことができる。	3.15	3.12	- 0.03	-0.01	
4	【書くこと】基本的な語彙，構文，よく使われる結束表現（接続詞，副詞等）を使って，つながりのある文章を書くことができる。	3.06	2.95	- 0.12	-0.04	
5	【話すこと】絵を見ながら，本文の内容を筋道を立てて分かりやすく話すことができる。（Summary）	2.95	2.93	- 0.02	-0.01	
6	【話すこと】パラグラフ構成を明確にした原稿を準備し，簡単なメモを見ながら，自分の意見や感想を，具体的な例を付け加え，筋道を立てて分かりやすく話すことができる。（プレゼンテーション）	2.71	2.93	0.22	0.08	**
7	【話すこと】個人で２分程度，比較的自然な言葉遣いで，伝えたいことをある程度的確に発表することができる。	2.60	2.75	0.15	0.06	
11	本文に出てきた単語は覚えた。	3.29	3.00	- 0.29	-0.09	**
14	本文についてのQuestionに答えることができる。	3.08	2.88	- 0.20	-0.06	*

は難しかったことを考えると，少なくとも単元CAN-DOリストに関しては，生徒の自己評価は上昇傾向にあり，「できる感」を持つようになったと言える。

しかし追加で加えていた他の項目に関しては下降傾向にあり，11と14で有意差が確認された。確かに教材は難しくなっているが，教えている側から見ると6月から9月にかけて下降したとは思えず，ここでも「自己調整学習」質問紙調査同様，生徒の自己評価が厳しくなってきている様子がうかがえる。

またCAN-DO Check Sheetの自由記述欄にはその授業内で「できたこと」を書くようにと指示を出したのだが書かせてみると「～ができなかった」という記述が多くみられた。再三「できなかったこと」ではなく「できたこと」を書くようにと伝えたが，「～が難しかった」「もっと～したい」「もっと～すべき」「～できたはず」と表現は変わったものの最後まで自分に厳しく内省する様子がうかがえた。

この結果から考えられることは，生徒は自己評価をするとき自分に厳しく過小評価しがちであるので，CAN-DOリストやCAN-DO Check Sheetは使い方を誤ると「できる感」ではなく「できない感」を持たせ，自己効力感を低下させるものになってしまう危険性があるということである。ではどうすればそれを避けられるのであろうか。有意に上昇した6や上昇傾向にある7はReflectionをする直前に単元の最終ゴールとして2分程度のプレゼンテーションを実施し，ほとんどの生徒がそれをクリアしていた。一方下降した11に関しては直前に単語テストをしその点数が生徒にわかっていた，14は単元の初期の活動でReflectionをする直前の活動ではなかったという特徴がある。つまり授業内の活動やタスクをCAN-DOリストやCAN-DO Check Sheetとうまく連動させ，実施した活動やタスクができた直後にその振り返りをさせ，生徒に「できた感」

を実感させる必要がある。またCAN-DOリストにはできるだけあいまいな表現は避け，生徒が実行可能な具体的数値目標等を入れていくことで，生徒が過度に自己評価を厳しくしたり過小評価したりすることを避けられると考えられる。

5.3 (RQ3) CAN-DOリストとCAN-DO Check Sheetの導入が，実際に英語力の向上につながるのか？

本校では年に1度12月にGTEC for STUDENTSを実施しており，3年生12月（生徒のレベルに合わせてCore，Basic，Advancedを受験）と4年生12月（生徒のレベルに合わせてBasic，Advancedを受験）のスコアを比較した。全体のスコアの平均は431.6から488.2へと上昇しており，GTECで測られる英語力は向上していることが明らかになった。しかし前期（4月～9月）にこの実践に参加した実験群の生徒が491.6点，その他の統制群の生徒が486.3点で，その間に大きな差はなかった。GTECは実践の期間とずれており他の要素も多く影響していたこと，他の客観的なテストを実施することができなかったことから，今回はCAN-DOリストとCAN-DO Check Sheetの導入と英語力の向上との直接的関連性は検証することができなかった。

5.4 (RQ4) Reflection Sheetの自己評価とCAN-DO Check Sheetの自由記述内容は生徒の英語力と相関や関連性があるのか？

後期は習熟度別クラスで授業を行っていたので，レベルの異なるそれぞれのクラスでCAN-DOリストとCheck Sheetの実践を行い，Reflection Sheetの質問項目1～7（図4参照）の集計結果がいかに異なるかを分析した（表10）。ただし習熟度別クラスは定期考査の結果と本人の希望を加味して分けているので完全に生徒の英語力が反映しているわけではない。また期末考査はIntermediate/AdvancedクラスとBasicクラスとで20点分が別問題であった。

■表10：3月のReflection Sheetの集計結果と期末考査やプレゼンテーション評価の平均点

Reflection Sheet 質問項目	1	2	3	4	5	6	7	期末考査（100点満点）	プレゼン（10点満点）
12組Basic（n=22）	3.05	3.29	3.00	2.76	3.10	3.00	2.80	43.2	7.98
3組Intermediate（n=17）	3.18	3.76	3.41	3.12	3.29	3.24	3.24	62.9	9.18
45組Advanced（n=23）	3.00	3.26	3.13	3.00	3.00	2.78	2.70	70.7	9.13

質問項目と期末考査の素点，プレゼンテーションの評価（JTEとALTによる）の相関分析を行うと，2,3,4と期末考査に，1,2,6とプレゼンテーションの評価に弱い相関は確認されたものの，必ずしもBasicクラスの自己評価が低いわけではない。むしろAdvancedクラスのほうが低い項目がある。この結果から，CAN-DOリストとCheck Sheetは様々な英語力の生徒に対して使い方次第で「できる感」を実感させることができることがわかった。

授業者として振り返ってみて，自己評価は本人の実際の英語力よりもむしろクラスの雰囲気に左右されている印象がある。Advancedクラスでは英語のできる生徒が多く集まっているため，他の生徒との相対評価をしたりより高い目標を目指したりするため自己評価は低くなる。3組Intermediateクラスは必ずしも英語力が高いわけではないが，単独HRクラスの生徒で構成されておりお互いに助け合いアットホームな雰囲気であった。その授業の雰囲気が各自の高い自己評価につながったのではないかと考えられる。

CAN-DO Check Sheetの自由記述内容は，4.2.7で述べたように後期にクラスによって使い方を変えたために，当初計画していたクラスによる比較検証をすることはできなくなったが，Basicクラスで「感想」「わからなかった点」「教師に伝えたいこと」など自由に書かせたところ，「～ができなかった」という表現はほとんど見られな

くなった。また授業者は授業直後にその記述を読むことで生徒の理解度や思いをすぐにキャッチして次の授業にフィードバックすることができたので，授業者と生徒の良い連絡シートとなった。Advancedクラスでは毎時間授業内で扱った内容について感想や意見などを英語で数行書くことがライティング力と思考力を高める良い機会となった。どのクラスにおいてもその教材を通して授業者が伝えたいと思っていることが伝わっているか，生徒はどう感じているかがよくわかり，授業者にとっても有益なツールとなった。

6 まとめ

本研究では第1にCAN-DOリストとCAN-DO Check Sheetの導入が自己調整学習の向上に効果があることがわかった。今回の実践では前半に自己調整学習の質問項目の平均値がいったん下がったために，1年間の伸びは確認することができなかったが，もし前半の下降がなければ，有意に上昇していたと予想される。今後は前半の下降の原因を追究し，そこで下降しない工夫を加えていきたいと考える。第2にCAN-DOリストやCAN-DO Check Sheetは使い方を誤ると「できる感」ではなく「できない感」を持たせ，自己効力感を低下させるものになってしまう危険性がある一方，使い方次第で様々な英語力の生徒に対して「できる感」を実感させることができることもわかった。

私自身が今回の実践中に行った授業改善で効果的であったと考えるのは、協同での活動を増やしたこと、活動でも振り返りでも生徒の自由度を増やしたこと、タスクと評価を連動させる工夫をしたこと、そこに他己評価も入れたことなどがあげられる。ただこれらが自己調整学習の向上にいかに寄与するかは今後検証が必要である。

最後に本研究から得られた結果をもとに，教育的示唆を以下のとおりまとめる。

①自己調整学習者を育成するためには，CAN-DOリストとCAN-DO Check Sheetのような意図的な介入が必要である。

高1で英語嫌いの生徒が急に増える現状において，CAN-DOリストとCAN-DO Check Sheetはそれを食い止める有効なツールになりうると考えられる。何の介入もしなければ生徒は自己効力感や内発的動機付けを低下させたままで英語学習を続けていくことになる。

②CAN-DOリストとCAN-DO Check Sheetを導入する際には，常に生徒の自己調整学習に関する項目，特に自己効力感や内発的動機付けの変化を確認する必要がある。

CAN-DOリストとCAN-DO Check Sheetは今後自律学習者を育成するツールとして広がる可能性があるが，不用意に導入するだけでは生徒に「できない感」を持たせる結果となってしまい，諸刃の剣になりかねない。CAN-DOリストを提示し高い目標を示した時，生徒は自己効力感を失う可能性があり，その時教師は素早くその変化に気づき，励ましや声掛けをするなど対応することが大切である。

③CAN-DOリストとCAN-DO Check Sheetは導入するだけでなく，授業との連動を意識した工夫や注意が必要である。

生徒がCAN-DOリストとCAN-DO Check Sheetを利用して自己効力感を維持するためには，CAN-DOリストと授業を連動させ，生徒がCAN-DO Check Sheetを記入するときに概ねできたと実感できるようにアクティビティやタスクを組み込んでいく必要がある。またCAN-DO Check Sheetはあいまいな表現を避け，できるだけ具体的で生徒が実行可能な数値目標等が含まれていることが望ましい。

④生徒の自己評価は教師の予想以上に低いので，自己効力感を失わせることがないよう配慮が必要である。

生徒は予想以上に自分に対して厳しく，低い自己評価をしがちであるが，それを続けると自己効力感の低下へとつながる。できることを実感させたうえで，自己評価をさせたい。また自己評価に客観性を持たせる工夫をしたり，他己評価も取り入れていくべきであろう。

⑤CAN-DOリストとCAN-DO Check Sheetは教師の授業改善の有効なツールであり，大いに活用すべきである。

常にCAN-DOリストを意識することで指導の長期短期の目標が明確になり，授業とCAN-

DOリストの連動を意識することで目標に到達するためのタスクを配列に注意しながら積み重ねていくようになった。またCAN-DO Check Sheetの生徒の自由記述は授業改善の参考になった。

謝辞

本研究を実施する貴重な機会を与えてくださいました，公益財団法人 日本英語検定協会の皆様，選考委員の先生方，特に助言担当の小池生夫先生に心より感謝申し上げます。

また実践協力校の生徒の皆様，御協力いただいた先生方，本当にありがとうございました。

参考文献（＊は引用文献）

＊ Hiromori, T. (2006). Gaikokugogakushusha no dokiduke wo takameru riron to jissen [Theory and Practice to increase foreign language learners' motivation]. Tokyo: Tagashuppan.

兵庫県教育委員会．(2015).『「兵庫版基本CAN-DOリスト」及びその活用について』.

＊ Maeda, H. (2002). Strategy use and achievement by Japanese high school EFL learners: Relationship with motivation and cognitive evaluation of learning. *Language Education & Technology,* 39, 137-148.

＊ 文部科学省．(2013).『各中・高等学校の外国語教育における「CAN-DO リスト」の形での学習到達目標設定のための手引き』.

＊ 文部科学省．(2013).「英語教育改革実施計画」.

＊ 文部科学省．(2016).「英語教育実施状況調査」.

＊ Mori, Y. (2004). Relations between self-efficacy and learning strategies of English in college students. *Japan Society for Educational Technology,* 28, 45-48.

＊ 長沼君主．(2008).「Can-do尺度はいかに英語教育を変革しうるかーCan-do研究の方向性」． ARCLE REVIEW. No.2, 50-77.

＊ 長沼君主・小川隆夫．(2010).「小学校英語教育におけるCan-Do尺度の開発とスキル指導の効果の検 証―スキルと情意のバランスを求めて」.『中央教育研究所研究報告』. No.73. 7-27.

長沼君主．(2017).「小学校英語Can-Do 評価尺度活用マニュアル解説～中学校英語教科書接続Can-Do リスト試案～」.『小中連携を目指した英語学習者の自己効力と自律性を促進する授業設計と評価』. 1-8.

＊ 中田賀之．(2001).「英語学習動機づけ研究：量的データと質的データ融合に向けての試みー大学1年生の意識調査」. 九州国際大学紀要第48号，117-138.

＊ 中田賀之．(2011).「自己調整とその他の学習要因」． 佐野・岡・遊住・金子（編）『第二言語習得　言語習得理論から脳科学まで』． 東京：大修館書店．

＊ 岡崎浩幸.(2014).「CAN-DOリストに基づいた英語授業に関する高校生の意識調査」． 富山大学人間　発達科学研究実践総合センター紀要教育実践研究8号，117-125.

＊ Oxford, R. (1990). Language learning strategies: What every teacher should know. Boston: Heinle & Heinle.

＊ Pintrich, P. R., & De Groot, E. V. (1990). Motivation and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology,* 82(1), 33-40.

＊ 辰野千壽 (1997).『学習方略の心理学』． 東京：図書文化

投野由紀夫．(2013).『CAN-DOリスト作成・活用　英語到達度指標CEFR-Jガイドブック』. 東京：大修館書店．

Tsuda A. (2011). Exploring Self-regulated Language Learning among Japanese Senior High School Students. 兵庫教育大学院修士論文.

＊ 渡部真喜子．(2015).「CAN-DOリストを活用した英語科授業の実践」. 東書Eネット第31回東書教育賞受賞者論文，18-35.

＊ Yashima, T. (2000). Orientation and Motivation in Foreign Language Learning: A Study of Japanese College Students. *JACET Bulletin,* 31, 121-133.

＊ Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology,* 81(3), 329-339.

第30回 研究助成
B 実践部門 ● 報告Ⅲ ● 英語能力向上をめざす教育実践

ルーブリックを用いて段階を踏んだ，継続的なスピーキング活動の実践

研究者：埼玉県／川口市立高等学校　教諭 鶴田 京子（申請時：川口市立県陽高等学校 教諭）
《研究助言者：和田 稔》

概要

本研究は「話すこと」の指導にルーブリックを活用して段階的に指導を行った実践報告である。「話すこと」を「発表」，「やりとり」の2つの領域に分け，2種類の質問を帯活動で継続的に扱った。本校の生徒の実情に合わせた評価項目，評価基準を設定しルーブリックを作成した。ルーブリックは生徒と共有し，毎時間評価項目から1つ焦点をあて教師は生徒に意識させるように支援を行い「話すこと」の活動を実施した。活動後に自己評価を行い，実技テストまで自己評価を継続した。

事前，事後で行ったスピーキング抵抗感尺度では有意な差は見られなかったが，取得級，項目別の分析では差が見られる部分があった。自己評価は回を重ねるごとに評価が上がっており，自由記述からも継続的な活動が生徒の自信になっていることがわかり，自己肯定感を高めることにつながる指導の可能性になると考える。

1 はじめに

平成28年度英語教育実施状況調査によるとパフォーマンステストを実施する学校は年々増えてきている。次期学習指導要領では4技能の1つである「話すこと」が「話すこと（発表）」，「話すこと（やりとり）」の2つの領域に分かれる。「発表」は一方向に対し，「やりとり」は応答が求められるため，「話すこと」の指導に際してそれぞれの目標に応じた指導が必要となってくる。また，評価においてもコミュニケーションの内容が異なり，評価するポイントが違ってくるため評価項目も重要となってくる。そこで，指導と評価の一体化が不可欠となる。

本実践は，進路多様校における「発表」，「やりとり」の継続的な実践とルーブリックを用いた自己評価活動の実践報告である。実践の結果，評価活動がどのように生徒に影響を与えたか，評価項目と方法に関して考察する。

2 実践の背景

2.1 「話すこと」

平成21年版学習指導要領では，「聞くこと」，「話すこと」，「読むこと」，そして「書くこと」の4技能で示されていた目標が，次期学習指導要領では4技能5領域となり「話すこと」が「話すこと（発表）」，「話すこと（やりとり）」の2つに分けられる。この背景にはCommon European Framework of Reference（略称：CEFR）が，SpeakingをInteractionとProductionに分けている影響がある。そのため今後は「発表」，「やりとり」ごとの目標設定と評価が必要になってくると考えられる。本実践ではスピーキング活動を「発表」から「やりとり」へと段階を踏んで導入し，それぞれの活動を関連付けて継続的に指導を行った。

2.2 スピーキングの評価

学校教育現場におけるスピーキングの評価は学習者の言語学習成果の評価という側面を持つ。そして，その成果は生徒の発話を通して評価されるため，パフォーマンス評価を行うことが妥当である。パフォーマンス評価は他の技能よりも採点に時間と手間がかかる。しかし，「個人の英語力や学習成果を測定し評定を確定する作業を，スピーキング指導と並行させようとするー見困難な実践も，綿密なカリキュラムおよび単元指導案の計画と，評価の対象や観点，基準の明確化によって，評価を容易かつ正確にすることもできる。」（冨田，2011）と述べており，学校現場において活用しやすい評価方法が必要となっている。本実践では，本校の実情に合わせた評価項目を作り，また生徒と共有して実施していくことで生徒の自己評価がどのように変化していくのかを考察した。

2.3 ルーブリック

ルーブリックはパフォーマンス課題を評価するための表である。学習到達度を示す評価基準は観点と尺度から構成されている。パフォーマンス評価をする際，評価の観点を記述文で示し，その観点をどの程度達成できたかを尺度（数値）で表す。Can-doが「〜できる」の二項対立の評価に対し，ルーブリックはより段階的に評価項目が設定される。「ルーブリックを教師と生徒が共有すれば，教師と生徒の間にめざすべき目標が具体的に設定され，その課題を達成させようと教師は支援をしたり，生徒は良い成績を取ろうと努力したりすると考えられる。」（小沢，2015）本実践ではルーブリックを事前に生徒に示し自己評価を行っていくことで目標を共有し，段階的に目標に焦点をあて支援を行った。

2.4 帯活動

松沢（2014）によると，帯活動は「短時間継続的に行う投げ込み活動」で，短い対話や簡単なゲーム等で授業のレディネスを形成するためのものである。英語の関心を高める働きがあるー方で，教科書とは切り離した活動になるため教科書を扱う時間が減ってしまう問題がある。今回の実践では，継続的に活動ができる点に着目し，スピーキングの定着を図ることに重きを置いた。

3 実践方法

3.1 目的

本実践の目的は，「話すこと」の活動における「発表」，「やりとり」の指導方法とルーブリックの活用法を探ることである。「発表」と「やりとり」の活動を関連づけ，それにルーブリックの自己評価活動を組み入れることでよりよい活動になることが期待できる。報告では英検の取得級に分け自己評価活動の推移，自己評価と教師による評価の相違点，活動による気づきに焦点をあてる。

3.2 参加者

本実践には筆者の勤務する埼玉県の公立高校3年生4クラスに属する157名が参加した。卒業後の進路は多様であり，外国語学習へのモチベーションも様々である。継続的な学習を苦手としている生徒が多く，知識の定着が課題となっている。学校の取り組みとして年に1回英検の全員受験を実施している。

生徒はコミュニケーション英語IIIを週に4時間全員が受講している。運用上，週のうち3時間と1時間で担当者を分けており，本実践では週3時間の授業で実践を行った。（平成29年8月末〜12月上旬）授業はクラス単位（約40名）で行われており，担当者1人で週3時間のコミュニケーション英語IIIを指導していた。また不定期でＡＬＴが授業に参加し，指導補助を行っていた。

3.3 パフォーマンスタスクの作成

本実践の参加者の英検の取得状況を見ると，3年生になった段階での3級の取得率が67％，準2級が22％と3級から準2級へのハードルが非常に高いと考えられた。英検のホームページに

掲載されている試験内容によると，英検3級のスピーキングの形式・課題が「受験者自身のことなど」に対し，準2級では「受験者の意見など」となる。また扱われる場面・題材も「身近なことに関する話題」から「日常生活の話題」へと広がる。入学時から授業の帯活動で話す活動は実施してきており，基本的なあいづちや質疑の練習を行い，パフォーマンステストも実施してきた。英検の取得率が示しているように，身の回りのことについては簡単な英語で話すことはできるが，話題の幅が広がりさらに自分の意見を話す活動になるとつまってしまうという課題が見えた。そこで準2級の二次試験で過去に使用された問題を活用し，「発表」と「やりとり」の活動のタスクを作成することにした。実践するにあたってなじみが薄い話題になってしまうと自分の意見を導き出すことに更に時間がかかってしまうため，8月の事前アンケートを実施した際にトピックの親密度調査も行った。その結果，以下の2つの質問をタスクとして扱うことにした。

タスク1 Do you think reading comic books is good for children?

タスク2 Do you think it is a good idea for schools to take their students to foreign countries on school trips?

「発表」の活動では質問に対して自分の意見，理由とその根拠を1分程度述べることをタスクに設定した。第1段階として自分の意見を考えさせ，発話させることを目標とした。この目標を達成させることで次に設定している「やりとり」の活動の際に，会話のはじまりが円滑になると考えた。第2段階として自分の意見を元にして，文の構造や全体の構成を考えさせ，一貫性のある発話を促すようにした。

次の「やりとり」の活動では質問に対して自分の考えを述べたあと，相手の考えを受けて関連した対話をペアで約2分間続けるタスクを設定した。パフォーマンステストではその場で決められたペアと対話をするように設定していたため，会話の最初でつまずいて沈黙となってしまわないように，「発表」の活動で実施した内容を取り入れることにした。また，1年生の時から帯活動で基本的な対話の練習は実施してきているため，定着に差はあるがペアでの基本的な質疑は学習済みである。今回のタスクでは問いに対して深い対話に持って行きたいと考えた。

3.4 ルーブリックの作成について

ルーブリックの評価項目および評価基準については，『HOPE中高生のための英語スピーキングテスト』（今井・吉田 2007）や『英語テスト作成ガイド』（小泉・印南・深澤 2017）を参考にし，本校の実情を踏まえて項目を作成した。さらに，英検準2級のCan-doリストに書かれている内容と過去2年間のスピーキング活動の実践から現在本校の生徒に必要とされる項目について考慮に入れた。また，帯活動の際に評価項目を意識して活動できるよう基準を可能な限り具体的に記述するようにした。さらに，生徒が評価の観点を理解でき自己評価しやすく，教師が時間をかけず簡単に評価できるような評価表の作成を目指し，できるだけ文を短くするようにした。

3.4.1「発表」のルーブリック作成

「発表」の活動では自分の意見の理由とその根拠を述べる内容に関する項目と，文と全体の構成に評価項目を分けた。「発表」から「やりとり」まで同じ質問をタスクとして使用するため「やりとり」の活動に移った際にペアで意見が重ならないように理由を2つ考えさせることにした。また，質問に対して深く考えさせるために理由の根拠を入れた。

本校の生徒の特長として単文のみの発話が多いことも課題であった。ただ発言すればいいのではなく，自分の考えを可能な限り正確に伝えられるようにさせたいと考えこの項目を入れた。全体の構成は英語の論理構成を身に付けさせ，最終的にはライティングにも対応できるようにさせたいと考えこの項目を入れた（表1参照）。

3.4.2「やりとり」のルーブリック作成

「やりとり」の活動では自分の考えを述べるだけでなく，「やりとり」として必要な要素とされるペアでのやりとりを意識して作成をした。そ

■表1:「発表」のルーブリック

	内容		構成		流ちょうさ
	理由	根拠	文	全体	
4	理由を2つ適切に述べられていた。	理由に対する根拠をすべて述べている。	まとまりのある発話をし、複雑な文(whenや関係詞など)を用いたこともあった。	論理立てて話ができている。	5秒以上長い沈黙がなく、スムーズに話し、適切なポーズがとれている。
3	理由を2つ述べられていた。	理由に対する根拠をすべて述べているが不完全である。	さまざまな文の種類を使って、まとまりのある発話をしていた。	不自然な部分もあるが論理立てて話ができている。	5秒以上長い沈黙がないが、話すスピードが遅めであったり、適切なポーズがない。
2	理由を1つ適切に述べられていた。	理由に対する根拠を一部述べている。	簡単な文で発話をしていた。	不自然な部分が多いが論理立てて話している。	長い沈黙が1回あったが、スムーズに話せていた。
1	理由を1つ述べられていた。	理由に対する根拠を一部述べているが、不完全である。	簡単な文と単語で発話していた。	不自然で論理立てて話せている部分がほとんどない。	長い沈黙が1回あり、スピードも遅めであった。
0	理由を言えなかった。	根拠が言えなかった。	単語のみの発話であった。	全く論理だっていない。	沈黙が何回もあり、スピードも遅めであった。

■表2:「やりとり」のルーブリック

	自分から発話		相手への応答		流ちょうさ
	自分の考え(理由)	相手への質問	繰り返しとあいづち	質問の応答	
4	適切に述べられていた。	相手の理由や応答に対して適切な質問ができていた。	適切な繰り返しとあいづちができていた。	相手の質問に対し、常に適切な応答ができていた。	5秒以上長い沈黙がなく、スムーズに話し、適切なポーズがとれている。
3	不自然な部分もあったが述べられていた。	相手の理由や応答に対して不自然な部分もあったが、質問ができていた。	不自然な部分もあったがほぼ適切な繰り返しとあいづちができていた。	相手の質問に対し、ほぼ適切に応答ができていた。	5秒以上長い沈黙がないが、話すスピードが遅めであったり、適切なポーズがない。
2	不自然な部分がほとんどであった。	相手の理由や応答に対して不自然な質問であった。	適切な繰り返し、または適切なあいづちができていた。	相手の質問に対して、不自然な応答があった。	長い沈黙が1回あったが、スムーズに話せていた。
1	単語のみで言っていた。	相手に伝わらない質問であった。	繰り返し、またはあいづちがほぼ適切にできていた。	相手の質問に対して、単語で回答していた。	長い沈黙が1回あり、スピードも遅めであった。
0	考えを言えなかった。	質問ができなかった。	どちらもできていなかった。	相手の質問に全く答えられない。	沈黙が何回もあり、スピードも遅めであった。

のため評価の項目も「自分からの発話」と「相手への応答」に評価基準を分けた。「自分の考え」は「発表」の活動ですでに経験済みの内容であるためできることを前提としている。しかし「発表」に対し「やりとり」は相手とともに行う活動なので生徒にかかる負荷が大きい。そのため自信を持って発言できる環境を最低限整えることが必要であると考え，この項目を入れた。もう1つの

評価基準はお互いの考えを述べた後の自由会話における発話である。相手の発話によって質問の内容が変わるため，質問の的確さと質問そのものの適切さを観点に入れる必要があった。

「相手への応答」にある「繰り返しとあいづち」は1年生の時から帯活動で短い対話活動を実施していたため，基本的な表現は学習済みである。ここでは相手の発話をきちんと理解した上で応答しているかを測るために「繰り返し」の項目を入れた。さらに，自由会話における質問の応答を評価項目に入れ，「やりとり」として成立していたかどうか測ることにした（表2参照）。

3.5 指導手順

週3時間のコミュニケーション英語IIIの授業の最初の約10分を帯活動として，「話すこと」の活動を実施した。「発表」，「やりとり」どちらもペアによる対話を基本とした。「発表」の活動を12回（6回×2テーマ）行い，その後「やりとり」の活動を12回実施した。各活動の終わりにパフォーマンステストを実施した。

「発表」，「やりとり」それぞれ第1回目の授業でルーブリックの評価項目について説明をし，パフォーマンステストでも同様に評価することを説明した。その後，毎時間帯活動でルーブリックの評価項目の1つに焦点をあて、その評価項目を意識してその日のペアワークに取り組むよう指導をした。

3.5.1「発表」活動

ペアになり1人1分間，質問に対する自分の考えと理由，その根拠を述べる活動を行った。聞き手を作ることで，相手がいることを意識して発話するような状況を設定した。ペアワークを始める前に，評価項目の中から1つその日に意識するべきポイントを説明した。それぞれの時間のポイントは，表3の指導手順に示すとおりである。ペアワークでは1人目がテーマについて話し，聞き手がコメントをした後，自己評価表を記入。その後役割を代え同じ流れを繰り返した。最後に質問を受けて通常の授業に入った。

3.5.2「やりとり」活動

「やりとり」の活動は「発表」活動の内容を活かせるように同じ質問を扱った。約2分間，ペアそれぞれが自分の考えを述べた後，考えに関連した自由会話を続けるタスクにした。パフォーマンステストは学校の実情から1回2時間分の授業でしか実施できないことを踏まえ，質問2つを交互に扱うことにした。「発表」の時と同様，活動を始める前にその日に意識するべきポイントを説明してから実施をした。（表4）自由会話では相手が代わると質問の内容も大きく変わることが想定されたため，すべての評価項目を扱った後は，毎時間座席を1つずつずらしペアが代わるように設定をした。

■表3:「発表」活動の指導手順

	ルーブリック評価項目	実施内容
1	内容―理由	ブレインストーミングをして理由を考える
2	内容―根拠	理由の根拠を考える
3	構成―全体	一貫性のある話し方を意識する
4	構成―文	相手に伝わる英語を考える
5	流ちょうさ	聞き手のことを考える
6	録音後の振り返り	録音した自分の発表を聞いて自己評価する

■表4:「やりとり」活動の指導手順

	ルーブリック評価項目	実施内容
1	自分の考えを述べる(タスク1, タスク2)	「発表」活動で述べた理由を振り返り, 相手に伝えやすい方法を考える
2	繰り返しとあいづち(タスク1)	相手の意見を聞いて簡単にまとめ, 適切なあいづちを考える
3	相手への質問と質問の応答(タスク1)	考えを聞いて質問を考える 質問を理解し適切な応答をする
4	繰り返しとあいづち(タスク2)	相手の意見を聞いて簡単にまとめ, 適切なあいづちを考える
5	相手への質問と質問の応答(タスク2)	考えを聞いて質問を考える 質問を理解し適切な応答をする
6	流ちょうさ(タスク1)	聞き手のことを考える
7	全体(タスク2)	いままで話したことのないペアと会話をし, 改善点を考える
8	全体(タスク1)	いままで話したことのないペアと会話をし, 改善点を考える
9	全体(タスク2)	4人グループになり自分たちの会話を 聞いてもらい改善点を考える
10	全体(タスク1又はタスク2)	パフォーマンステスト後自己評価する

3.6 パフォーマンステストについて

「発表」,「やりとり」それぞれの活動後, パフォーマンステストを実施した。「発表」は本校にあるＬＬ設備を利用し, 全員が一斉に録音, 保存をして提出させた。一方「やりとり」は授業時間2時間を使用して別室で試験を実施した。ペアはくじ引きによるランダム形式で, テストの終わった生徒がくじを引き次に行う生徒を決定するというシステムにした。

評価は「発表」は録音されたものを聞いて評価し,「やりとり」は学校の実情を踏まえ, その場で評価を行った。念のため, すべてのテストは録画を行った。

4 検証方法と結果

4.1 ルーブリックによる生徒の自己評価の変化

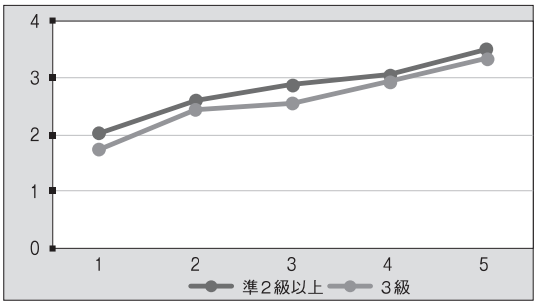
今回のパフォーマンスタスクは英検準2級の問題を使用したため, 準2級の取得の有無を基準として検証を行った。ここでは準2級以上の取得者と3級の取得者の2つのグループを作成し比較を行った。

4.1.1「発表」ータスク1　Do you think reading comic books is good for children?

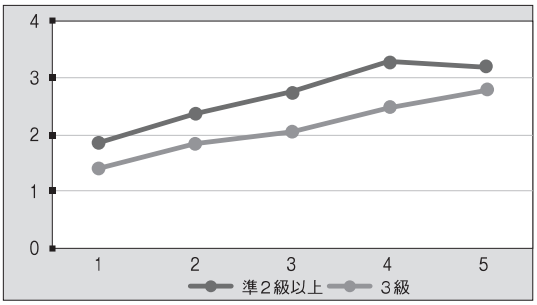
第1回目の活動の時にブレインストーミングを実施し, 意見を考える活動のみで終わってしまったため, 実際に発表活動を行った2回目から6回分の自己評価の記録となっている。図1-1～1-5のグラフの傾きから自己評価の数値が段階的に上がっていることがわかる。

■表5:「発表」タスク1における自己評価表の変化（準2級以上 N=38，3級 N=69）

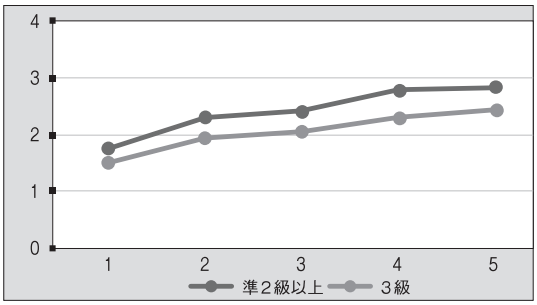
		2回目		3回目		4回目		5回目		6回目	
	取得級	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
理由	準2級以上	2.03	1.00	2.61	0.86	2.88	0.83	3.06	0.76	3.50	0.55
	3級	1.75	1.06	2.46	1.02	2.57	1.14	2.94	1.00	3.36	0.59
根拠	準2級以上	1.86	1.03	2.36	1.03	2.74	0.85	3.24	0.84	3.21	0.80
	3級	1.42	1.14	1.83	1.08	2.05	1.10	2.46	1.05	2.77	1.17
文の構造	準2級以上	1.75	0.89	2.28	0.73	2.41	0.65	2.76	0.69	2.82	0.76
	3級	1.52	0.88	1.91	0.84	2.02	0.85	2.26	0.86	2.43	0.63
全体の構造	準2級以上	2.03	1.06	2.42	0.95	2.65	0.72	3.06	0.91	2.97	0.71
	3級	1.52	0.88	1.91	0.84	2.11	0.96	2.43	1.02	2.65	0.97
流ちょうさ	準2級以上	1.75	1.30	2.44	1.01	2.76	0.91	3.06	1.03	3.18	0.76
	3級	1.67	1.17	2.14	1.09	2.32	0.99	2.58	1.12	2.93	0.90



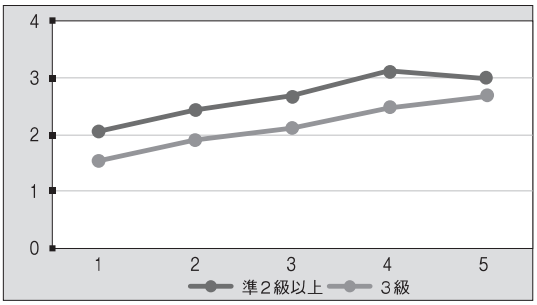
■図1-1: タスク1　理由



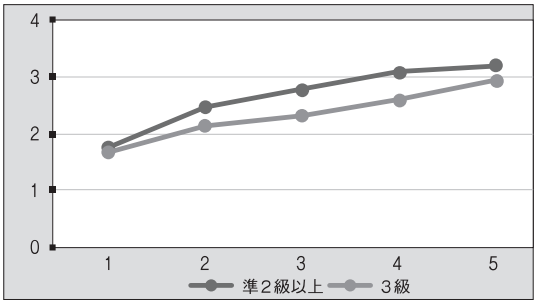
■図1-2: タスク1　根拠



■図1-3: タスク1　文の構成



■図1-4: タスク1　全体の構成



■図1-5: タスク1　流ちょうさ

4.1.2「発表」—タスク2　Do you think it is a good idea for schools to take their students to foreign countries on school trips?

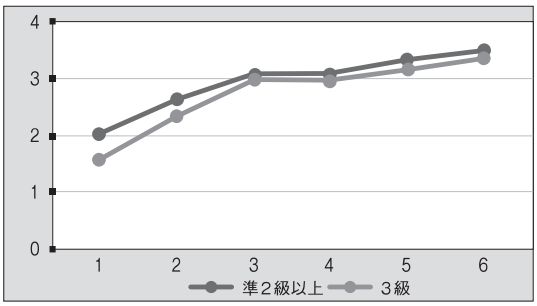
タスク2はタスク1と質問は異なるものの指導手順は同じであったため，表6は第1回目から自己評価活動を開始した計6回分の自己評価の記録となっている。タスク1と同様に回を増すごとに自己評価の得点が高くなっている。(図2-1〜2-5)「理由」に関しては回数を重ねることで自分の決まったパターンができあがり点数が安定していた。

4.1.3「やりとり」活動

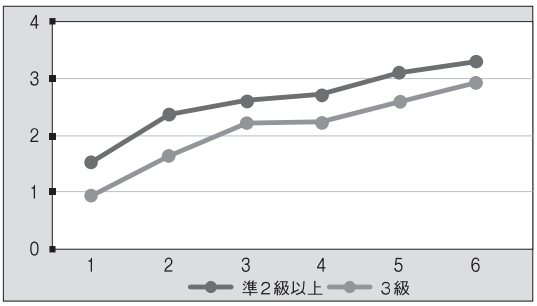
「やりとり」は「発表」で使用したタスク2つを交互に扱ったため，扱われたタスクの順番や項目によって自己評価に凸凹が見られる。「自分の考え」は「発表」で考える活動を行っていたため，級を問わず全体を通して点数が高くなっている。3回目の自己評価は準2級以上取得者と3級取得者の差があまり見られないだけでなく，「相手への質問」(図3－2)に関しては3級の生徒の方が自己評価の平均が高くなっている。この回は「相手への質問と質問の応答」に焦点を当てた回であり，相手の考えを聞いてそれに対応する質問や応答がどの生徒にとっても難しかったと考えられる。またペアで会話を始める前に相手の考えに即した質問と応答をするように助言をしたため，その解釈で評価が分かれた。そして，6回目の自己評価の項目は「流ちょうさ」を焦点に当てており，タスクがその前の時間に実施したものと変わったため，とまどってしまい自己評価が下がった。7回目以降はペアが毎時間変わり，相手によって点数の上下が見られる生徒もいたが，全体としては回を重ねるごとに平均が上がっていった。

■表6:「発表」タスク2の自己評価表の変化(準2級以上 N=38，3級 N=69)

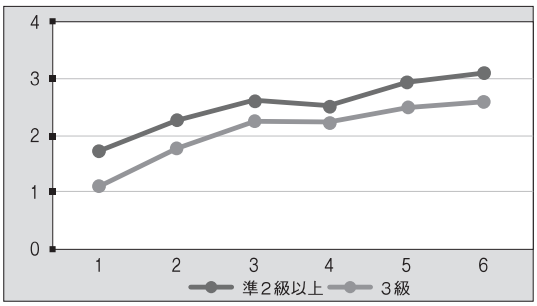
		1回目		2回目		3回目		4回目		5回目		6回目	
	取得級	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
理由	準2級以上	2.06	1.17	2.65	0.97	3.05	0.80	3.08	0.97	3.33	0.65	3.49	0.55
	3級	1.57	1.33	2.32	1.05	2.97	0.82	3.02	0.82	3.16	0.73	3.38	0.64
根拠	準2級以上	1.54	1.13	2.38	0.97	2.65	0.91	2.76	1.08	3.13	0.76	3.32	0.62
	3級	0.94	1.12	1.65	1.08	2.23	1.05	2.28	1.07	2.63	0.99	2.97	0.87
文の構造	準2級以上	1.71	0.94	2.24	0.84	2.59	0.88	2.54	0.92	2.97	0.71	3.11	0.69
	3級	1.11	1.03	1.77	0.80	2.24	0.78	2.25	0.72	2.50	0.74	2.59	0.71
全体の構造	準2級以上	1.74	1.05	2.35	1.00	2.73	0.89	2.70	1.06	3.13	0.72	3.35	0.62
	3級	1.16	1.09	1.85	1.10	2.27	0.94	2.30	0.95	2.52	1.03	2.85	0.86
流ちょうさ	準2級以上	1.83	1.28	2.44	1.12	2.95	1.01	2.92	1.15	3.33	0.83	3.35	0.71
	3級	1.25	1.23	2.13	1.27	2.59	1.14	2.63	0.87	2.95	0.87	3.10	0.75



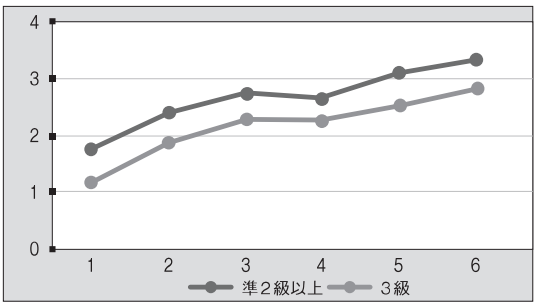
■図2-1: タスク2　理由



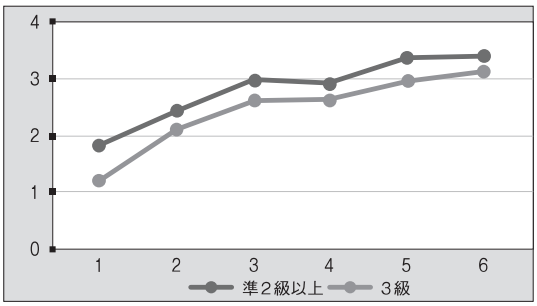
■図2-2: タスク2　根拠



■図2-3: タスク2　文の構成



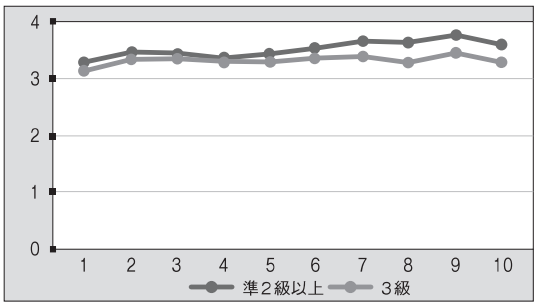
■図2-4: タスク2　全体の構成



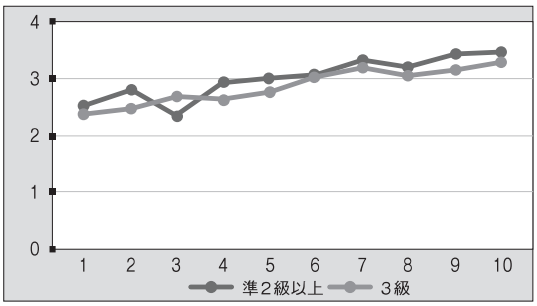
■図2-5: タスク2　流ちょうさ

■表7:「やりとり」の自己評価の変化（準2級以上 N=36，3級 N=67）

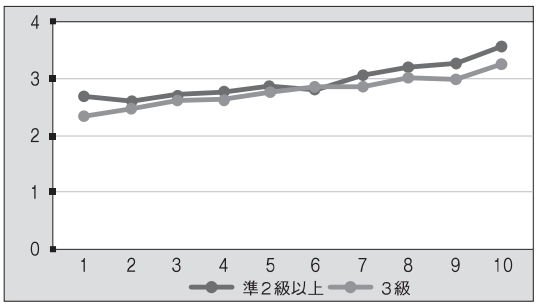
		1		2		3		4		5		6		7		8		9		10	
	取得級	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
自分の考え	準2級以上	3.28	0.93	3.47	0.65	3.40	0.80	3.33	0.72	3.44	0.85	3.53	0.61	3.65	0.72	3.63	0.59	3.77	0.48	3.60	0.49
	3級	3.12	0.85	3.33	0.70	3.35	0.73	3.29	0.67	3.31	0.58	3.35	0.71	3.39	0.72	3.28	0.87	3.45	0.77	3.28	0.75
相手への質問	準2級以上	2.56	1.17	2.79	1.11	2.34	1.43	2.94	1.01	3.00	1.00	3.06	0.93	3.32	0.67	3.20	0.82	3.46	0.81	3.46	0.97
	3級	2.37	1.28	2.48	1.26	2.69	1.16	2.63	1.14	2.77	1.00	3.04	0.82	3.20	0.77	3.04	0.96	3.15	0.86	3.30	0.91
繰り返しとあいづち	準2級以上	2.69	1.02	2.62	1.00	2.74	0.91	2.79	0.91	2.88	0.83	2.81	0.92	3.06	1.03	3.20	0.75	3.29	0.74	3.60	0.49
	3級	2.32	1.08	2.51	0.95	2.65	0.98	2.63	0.89	2.77	0.80	2.89	0.88	2.87	0.92	3.03	0.89	3.12	0.90	3.27	0.89
質問の応答	準2級以上	2.58	1.04	2.71	1.07	2.65	1.11	2.88	0.98	2.97	0.92	2.91	0.72	3.21	0.72	3.26	0.69	3.37	0.64	3.60	0.55
	3級	2.36	1.10	2.60	0.83	2.61	1.09	2.55	0.93	2.72	0.83	2.84	0.82	2.87	0.90	3.06	0.88	3.00	0.93	3.27	0.87
流ちょうさ	準2級以上	2.67	0.97	2.85	0.94	2.86	0.93	2.88	0.91	3.03	0.89	2.94	0.97	3.21	0.72	3.14	0.59	3.29	0.88	3.49	0.55
	3級	2.44	1.03	2.70	0.81	2.72	0.92	2.74	0.85	2.94	0.74	3.00	0.71	2.84	0.87	2.97	0.84	2.97	0.84	3.31	0.85



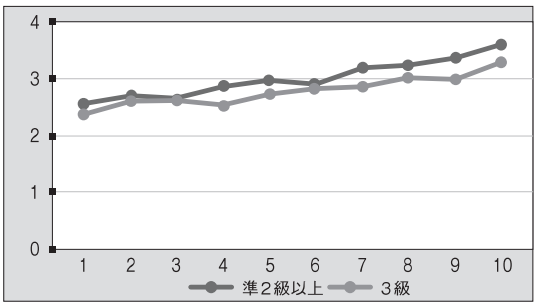
■図3-1: やりとり　自分の考え



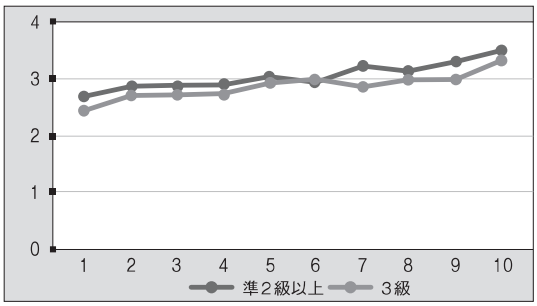
■図3-2: やりとり　相手への質問



■図3-3: やりとり　繰り返しとあいづち



■図3-4: やりとり　質問の応答



■図3-5: やりとり　流ちょうさ

4.2 自己評価力について

表8は,「発表」のパフォーマンステスト後の自己評価と教師の評価のずれの度合いを示したものである。評価のずれが大きいことがわかる。さらに細かく見ていくため, それぞれの評価項目ごとの一致の割合と, 不一致の割合を調べた。特に不一致の割合に関しては, 教師と生徒どちらの評価が高くて不一致になっているのかを明らかにした。(表9)それによると多くの項目で教師の評

■表8: 自己評価と教師の評価のずれ(N=38, N=69)

	1回目		2回目	
	準2級以上	3級	準2級以上	3級
ずれなし	0	0	2	0
1つずれ	5	9	10	8
2つずれ	8	19	5	14
3つずれ	9	16	10	21
4つずれ	13	17	7	17
5つずれ	3	8	4	9

価が高くつけられており，生徒が自己評価する際に自分の評価を低くつけてしまっている傾向にある。

■表9：評価項目別評価の一致，不一致の割合

1回目	理由		根拠		文構造		全体構造		流ちょうさ	
	準2級以上	3級	準2級以上	3級	準2級以上	3級	準2級以上	3級	準2級以上	3級
教師の評価が高い	31.6%	26.1%	36.8%	33.3%	63.2%	53.6%	55.3%	39.1%	44.7%	50.7%
生徒の評価が高い	18.4%	27.5%	18.4%	27.5%	5.3%	4.3%	15.8%	21.7%	13.2%	10.1%
一般	50.0%	46.4%	44.7%	39.1%	31.6%	42.0%	28.9%	39.1%	42.1%	39.1%

2回目	理由		根拠		文構造		全体構造		流ちょうさ	
	準2級以上	3級	準2級以上	3級	準2級以上	3級	準2級以上	3級	準2級以上	3級
教師の評価が高い	39.5%	23.2%	28.9%	21.7%	39.5%	47.8%	31.6%	40.6%	39.5%	44.9%
生徒の評価が高い	2.6%	29.0%	21.1%	34.8%	10.5%	15.9%	26.3%	30.4%	18.4%	18.8%
一般	57.9%	47.8%	50.0%	43.5%	50.0%	36.2%	42.1%	29.0%	42.1%	36.2%

■表10：自己評価と教師の評価のずれ（N=36，N=67）

	準２級以上	3級
ずれなし	2	6
1つずれ	11	10
2つずれ	7	14
3つずれ	3	11
4つずれ	11	12
5つずれ	2	14

同様に「やりとり」に関しても自己評価と教師の評価のずれを見ていくと，「発表」同様に評価にずれが生じていることがわかる。項目ごとに見ていくと，「発表」同様，教師の方が高く評価をつけていることが多いことがわかる。

■表11：評価項目別評価の一致，不一致の割合

1回目	理由		根拠		文構造		全体構造		流ちょうさ	
	準2級以上	3級	準2級以上	3級	準2級以上	3級	準2級以上	3級	準2級以上	3級
教師の評価が高い	36.1%	47.8%	30.6%	32.8%	38.9%	41.8%	33.3%	28.4%	38.9%	31.3%
生徒の評価が高い	8.3%	11.9%	5.6%	17.9%	27.8%	22.4%	11.1%	19.4%	13.9%	28.4%
一般	55.6%	40.3%	63.9%	49.3%	33.3%	35.8%	55.6%	52.2%	47.2%	40.3%

4.3 アンケート

磯田（2007）のスピーキング抵抗感尺度を事前，事後で行った。（資料1）これは段階的なスピーキング活動とルーブリックで評価の観点を明確にすることでスピーキングに対する意識に変化が見られるか調査するためである。また，すべての活動が終わった最後の授業で記述式のアンケートを併せて実施した。

4.3.1　スピーキング抵抗感尺度

参加者157名のうち回答があった148名を分析の対象とした。スピーキング抵抗感尺度を３つの要因（能力・不安・回避）に対して事前・事後で分散分析を行った。これに関してはどの項目も有意差は見られなかった。（表１２）

■表12: 事前・事後の分散分析

要因		*SS*	*df*	*MS*	*F*	*p*
抵抗感能力	グループ間	24.986	1	24.986	1.204	.273
	グループ内	6101.500	294	20.753		
	合計	6126.486	295			
抵抗感不安	グループ間	.03	1	.03	.001	.973
	グループ内	7785.885	294	26.483		
	合計	7785.916	295			
抵抗感回避	グループ間	50.248	1	50.284	1.942	.164
	グループ内	7612.257	294	25.892		
	合計	7662.541	295			

次に取得級による抵抗感の差を検証した。ここでは取得級と３つの要因による一元分散分析を行った。（表13）その結果，抵抗感尺度の能力１（（F (3,295) =3,368,p<.05）と回避（F (3,295) =9.609 p<.001）に関して有意差が見られた。この２要因に関して多重比較を行った。能力に関しては，「取得級のない生徒」と「２級取得の生徒」の間で有意差（p<.05）が見られた。また，回避に関しては「取得級のない生徒・3級取得の生徒」と「２級・準２級の生徒」の間に有意差（p<.01）が見られた。

■表13: 取得級別の分散分析

要因		*SS*	*df*	*MS*	*F*	*p*
抵抗感能力	グループ間	204.905	3	68.302	3.368	.019
	グループ内	5921.581	292	20.279		
	合計	6126.486	295			
抵抗感不安	グループ間	130.919	3	43.640	1.665	.175
	グループ内	7654.997	292	26.216		
	合計	7785.916	295			
抵抗感回避	グループ間	688.481	3	229.494	9.609	.000
	グループ内	6974.060	292	23.884		
	合計	7662.541	295			

4.3.2 アンケートの記述文より

自由記述では①帯活動で自己評価を行うことに対する感想や意見，②同じ質問で帯活動を継続的に行うことに対する感想や意見，③「発表」「やりとり」の活動に関しての感想や意見に分けてまとめていく。

4.3.2.1 自己評価活動に関して

自己評価活動に関しては，以下のような肯定的なコメントがあった。（原文ママ）

・自分ができない所をみなおす機会になった
・自分の苦手や出来ないことが明確に出ていたので意識して練習できました
・自分で振り返ることによって自分に足りないところがすぐにわかるので，活用していくべきだと思う

一方で，すでに準2級を取得している生徒の中には「自分が今どれくらいできるのかわかるけど正直めんどう」とコメントを書いた生徒もいた。

4.3.2.2 同じ質問で継続的に行うこと

継続的に行うことで自分の話す内容を深めることができたといったコメントや，継続的に行うことでモチベーションを高めることにつながった，など肯定的なコメントが見られた。また，全体を振り返って自己反省をするコメントや様々なトピックを扱ってほしいという要望のコメントから同じ質問を扱うことへの課題も見えた。さらに長期的に帯活動を実施する際にはこの点にも留意する必要がある。以下コメントを紹介する。（原文ママ）

・同じトピックを続けて練習することによりどんどん濃い内容の話をすることができるようになる
・同じトピックで練習することで，深く考えることができました
・同じトピックなので、自分がどの程度上がったかを実感出来るので，すごくやりやすいと思いました
・同じトピックで練習を続ける方が改善しなければいけないところが毎回よくわかる
・毎回同じ事を言うからやればやるだけできるようになってたくさん練習しようと思った
・繰り返すことで頭にも入ってきてやりやすかった。
・慣れるとさらに発展させてみようか等と考えて行動に移したりできるので良い
・何度も練習することで長く会話を続けられたので良かったです
・5回あったけど、3回目くらいからあまり変わらなかった気もした。工夫をすればよかった
・途中で少しあきてしまったかも・・・。練習にはなった
・色んなのをやりたい

4.3.2.3「発表」，「やりとり」について

２つの活動を通して「発表」と「やりとり」で同じ質問を扱ったことから，「やりとり」の導入がスムーズにいったとのコメントがあった。これは取得級の低い生徒からコメントで少しでも自信を持って次の活動に参加することができたことがうかがえる。特に「やりとり」に関しては，相手の考えで話す内容を変えなくてはいけないため，応答に難しさを感じるとコメントした生徒が多かった。その一方で，毎時間異なるペアと会話を行ったことが深く考えるきっかけとなったり，新たな表現や視点を学ぶことができたなど，前向きなコメントも見られた。以下原文ママでコメントの一部を紹介する。

・理由は答えられても根拠はうまく答えられなかった。
・流ちょうさはなかなかできなかった。
・1分間スピーチより難しい。質問を考えたりあいづちをしたりとすることが多いからだと思う
・あいづちが伸びづらかった。もっとレパートリーを増やすべきだった
・1分間スピーチを先にしていたから「やりとり」は最初から少しできていた
・いろんな人と会話をすることでいろいろな質問を考えたり、どうやればスムーズにできるかが分かってよいと思いました
・たくさんの人とやることで、様々なやりとりができて自分の自信につながった。相手のいい所をまねて次に役立つこともあった
・たくさんの質問をきける。英語が得意な人苦手な人、仲の良い人そんなに良くない人でやりやすさが違うし、本番に誰とやるかわからないのでよかったとおもった
・自分が思いつかない質問の内容などを教えてもらえてよかった
・人によって内容も質問もちがったので考えることは大変だったけど、こういう質問もあるんだなとわかったりできました。
・たくさんの人のパターンを練習することでいろんな応答が練習できました
・内容が毎回違って考えることが多くためになった

・同じことを続けると内容はよくなったりスラスラ発言できるようになるけど、同時に同じ質問ばかりになってしまっているのではないかと思った
・毎回相手がちがうと質問なども違うので新鮮だなと思った

5 考察

本実践により，生徒の自己評価活動がどのような影響を与えたか，また評価項目とその実施方法に関して考察をする。今回はタスクを準2級レベルの会話に設定をしたこともあり，3級と準2級を合格している生徒の間には自己評価の記録に一定の差があった。しかしどちらのグラフも回を増すごとに自己評価の数値があがっており，級に関係なく自己評価活動を行うことで最終目標に向けて向上するように意識して取り組んでいる傾向が見られた。これは自由記述から自己評価活動を行うことでセルフモニタリングできる，と効果を感じている生徒がいることからも目標に向けた意識付けができたと言える。

次に，同じ質問を使用して，評価項目の目標に応じて継続的，段階的に帯活動で実施したことも，自由記述等で一定の効果を感じている生徒が多かったことから，意味のある活動であったと言える。特に「やりとり」でペアを変えて実施したことは，困難さを感じながらも新たな表現や視点を相手から学びを得ていたようである。しかし，毎時間自己評価を行うことを負担に感じている生徒もおり，自己評価の運用に関しては改善や工夫が必要である。また，今回は本校の実情と課題を基に評価項目を設定したが，項目によって自己評価の付け方に差が見られたことや生徒と教師の評価間にも差が見られたことから，具体的なモデルを事前に示す等，生徒が評価する内容をよりイメージしやすくするべきであったと考えている。

事前・事後に行ったスピーキング抵抗感尺度で有意差が見られなかったことから，今回の実践はスピーキングの抵抗感に影響を与えなかったと考えられる。しかし，3級取得者と準2級以上取得者の間には英語学習や英語会話に対する回避の感情に差が見られたことから，この感情に変化を与えられるような自己評価や活動の工夫をすることで，変化を生みだすことが可能となるのではないかと考えられる。この点に関してはまた違った観点からの検証が必要である。

6 まとめと今後の課題

本実践を通して，帯活動で継続的に話すことの活動を実施することは「発表」，「やりとり」どちらにおいても「話すこと」に一定の効果があったと考えている。特に進路多様校において，生徒は回数を重ねることで「話すこと」に慣れ，自信を持って話すことができるようになる。また，ルーブリックを活用することで生徒は目標を明確にすることができ，教師は生徒に段階的に支援することが可能となる。継続的かつ段階的な指導は，自己肯定感を高めることにつながる活動になると考えている。

今後の課題として，ルーブリックの活用のさらなる実践と検証である。生徒が抱えている課題や複数の評価者による一貫性，時間的制約など現場が抱える実情は様々である。多くの実践と検証が広がることで汎用性の高いルーブリックができあがっていくのではないかと考えており，今後も研究を継続していきたい。

謝辞

本研究の機会を与えてくださった公益財団法人 日本英語検定協会の皆さま，選考委員の先生方，とりわけ丁寧なご指導・ご助言をいただきました明海大学名誉教授の和田稔先生に深く感謝申し上げます。また，分析に協力いただいた三室戸元光先生，実践に協力していただいた申請時の勤務校であった川口市立県陽高等学校の生徒のみなさんにはこの場を借りて厚くお礼申し上げます。

参考文献（＊は引用文献）

石田祥一・西田正・斉田智里　（編）(2011).『英語教育学大系　第13巻　テスティングと評価』[第2版] 東京：大修館書店
泉惠美子・門田修平　編著　『英語スピーキング指導ハンドブック』. (2016).　東京：大修館書店
＊磯田貴道.　(2007).「英語スピーキング抵抗感尺度の作成」. https://ir.lib.hiroshima-u.ac.jp/files/public/2/24578/20141016145529381041/h-gaikokugokenkyu_11_41.pdf
今井裕之・吉田達弘　編著　(2007).『HOPE　中高生のための英語スピーキングテスト』東京：教育出版
旺文社. (2017).『2017年度版英検準2級過去6回全問題集』. 東京. 旺文社.
＊小沢浩. (2015).　「スピーキング分野における「英検Can-doリスト」活用の工夫―ルーブリックの活用を通して」STEP BULLITEN, vol.27, 184-216.
小塩真司. (2016).　『第2版　研究事例で学ぶSPSSとAmosによる心理・調査データ解析』.　東京：東京図書
小泉利恵・印南洋・深澤真（編）(2017).『実例でわかる　英語テスト作成ガイド』. 東京：大修館書店
寺内正典（編）(2012).『英語教育学の実証的研究法入門』　東京：研究社
＊冨田かおる・小栗裕子・河内千栄子　（編）(2011).『英語教育学大系　第9巻　リスニングとスピーキングの理論と実践』[第2版]　東京：大修館書店
松井市子. (2016).「評価rubricを活用した英語ライティング力と自己評価力の育成をねらった実践」. STEP BULLITEN, vol.27, 146-164.
＊松沢伸二. (2014).「帯活動―その正体と魅力」. TEACHING ENGLISH NOW, Vol 27, 2-3111.　東京：三省堂.
望月昭彦・深澤真・印南洋・小泉利恵　（編）(2015).『英語4技能評価の理論と実践ーCAN-DO・観点別評価から技能統合的活動の評価まで』. 東京：大修館書店

資料1：スピーキング抵抗感尺度

1. 私は、英語を話す力がないと思います。（能力1）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

2. 私は、人と英語で話す時、緊張します。（不安1）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

3. 私は、できれば人と英語で話したくありません。（回避1）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

4. 私が話す英語は、相手に意味が伝わらないと思います。（能力2）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

5. 私は、人と英語で話すことは恥ずかしいです。（不安2）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

資料1： スピーキング抵抗感尺度

6. 私は、人と英語で話すことは避けたいです。（回避2）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

7. 私は、英語で話してコミュニケーションできる力があると思います。（能力3）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

8. 私は、人と英語で話す時、どきどきします。（不安3）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

9. 私は、人とたくさん英語で話したいです。（回避3）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

10. 私は、今の英語力では英語で話すことはできないと思います。（能力4）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

11. 私は、人と英語で話す時、リラックスしています。（不安4）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

12. 私は、英語で話さなければならない時は、できるだけしゃべらないようにしていたいです。（回避4）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

13. 私は、英語で話して、自分で考えていることを相手に伝えることができないと思います。（能力5）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

14. 私は、人と英語で話す時、心配な気持ちになります。（不安5）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

15. 私は、英語で話さなければならない時でも、日本語で話したいと思うことがよくあります。（回避5）
①全くあてはまらない　②あまり当てはまらない　③やや当てはまらない　④どちらともいえない
⑤やや当てはまる　⑥おおむね当てはまる　⑦よく当てはまる

第30回 研究助成
B 実践部門 ● 報告Ⅳ ● 英語能力向上をめざす教育実践

高校生を対象にした「書くこと」に焦点をあてた協働学習が英文エッセイにおけるCEFRレベル別語彙使用に与える効果

研究者：愛媛県／愛媛県立西条農業高等学校　教諭 三好 徹明（申請時：愛媛大学附属高等学校 教諭）
《研究助言者：池田 央》

概要

本実践研究は，協働学習の一形態であるピア・レスポンスによる相互評価活動とそのあとに行ったリライト活動が英文エッセイにおけるCEFRレベル別語彙使用に与える効果の定量的分析を目的とする。筆者は「国際理解」をテーマに設定した英文エッセイの作成を夏休み課題（プレタスク）と冬休み課題（ポストタスク）として与え，それぞれのプロダクトを両方とも完成させた90名の日本人高校生に電子データで提出させた。ポストタスクの英文エッセイについては，プレタスクの英文エッセイをリライトしたものを提出させた。プレタスクとポストタスクにおける英文エッセイのプロダクトの使用語彙を，English Profileで公開されているText Inspectorによって，CEFRの6レベル（A1～C2）の語とレベル付けされていない語（Unlisted）に分類した。そのあと，プレタスクとポストタスクの英文エッセイにおけるCEFR　A1からC2までの語彙数に関して，上位・中位・下位の成績群別に分析を行った。結果として，ピア・レスポンスとリライト活動のあとでは，CEFR　Aレベル，特にCEFR　A2レベルの語彙が有意に増えており，成績上位群・中位群・下位群のいずれの参加者も読み手に分かりやすい単語を使用したことが示唆された。

1 はじめに

1.1 問題の所在（L2ライティング研究）

これまでの第二言語習得（Second Language Acquisition: SLA）における国内外のL2ライティング研究に関してふれながら，現在の日本の英語教育におけるL2ライティング指導の現状についてまとめておく。

Matsuda（2003）によれば，L2ライティング研究の歴史に関する解説は1990年代に始まったと言われるが，L2ライティングの教育的アプローチ研究そのものは既に1960年代に始まったというのがSLAの歴史において重要であり，多くの研究者が認めるところであると述べている。小室（編）（2001）によると，1970年代から1980年代にかけて米国においては，認知心理学や心理言語学の分野で，母語であるL1ライティングのプロセスに関する実証的研究が盛んに行われた。そのことによって，プロセス重視の研究に転換したことが，ESLライティングの指導に大きな変化をもたらし，その結果を受けて，大学におけるL2ライティングのプロセス研究も注目されるようになった（小室（編），2001）。

しかし，その当時の日本国内において，大学ではそれほどプロセス研究が進んでいるとは言えず，また，中学校や高等学校のライティング指導では，構文の暗記や和文英訳が依然としてなさ

れていた（小室（編），2001）。現在でもライティング指導は，英語教育における4技能の中でもっとも指導が遅れている分野であると指摘されている（馬場，2010）。その問題点を解消するため，2009年度より高校においては「英語表現Ⅰ」「英語表現Ⅱ」という科目が新たに学習指導要領（平成21年3月告示）に基づいて設定された。これらの科目については，「英語を通じて，積極的にコミュニケーションを図ろうとする態度を育成するとともに，事実や意見などを多様な観点から考察し，論理の展開や表現の方法を工夫しながら伝える能力」を養い，そしてその能力を伸ばすことが目標であった（cf. 文部科学省,2009, pp.88-90）。

学習指導要領が改訂されたあと，中学校や高等学校の指導者の取組によって，ライティング指導は改善されてきてはいるものの，依然として現場では，文法事項の理解・定着のための単文レベルでの英作文指導や多くても自由英作文にとどまっているのが現状であると指摘されている（馬場，2010）。日本国内の高等学校や中等教育学校を対象とした文部科学省（2016）の英語力調査によると，4技能全てのうち、特に「書くこと」「話すこと」について課題があるとの指摘があった。その指摘を受けて，近年，生徒や学生のライティングスキルを伸ばすために，英文エッセイ指導を取り入れ始めている高校や大学なども増えてきている。

ここ近年のSLA研究について概観すると，コーパス言語学で利用されている学習者コーパスを利用して，第二言語習得の仕組みを明らかにしようとする研究が散見されるようになった。例えば，Hawkins and Filipović（2012）は，CEFR（Council of Europe,2001）の言語尺度を用いて，コーパスから学習者のCEFRレベルを特定する基準特性（criterial feature）の抽出を試みようとしている。この基準特性とは，「熟達度を特徴付け，下位レベルから上位レベルを区別する学習者言語の特性」のことである（村上，2013,p.147）。CEFR基準特性に関連した研究として，平林（2016）が日本人上級英語学習者の自由英作文の語彙の豊かさを調査し，B2とC1レベルの学習者のCEFRに関連づけられた英語力熟達度と語彙の豊かさ（lexical richness）との関連を検証した。しかし，この研究においては，A1・A2・B1レベルの学習者の英語力熟達度と語彙の豊かさとの関連については，検証されてはおらず，今後，これらの同レベルの生徒や学生が書いたエッセイのテキストデータに基づいた検証も必要である。以上のような現状をふまえ，今回の研究では，4技能のうちライティングに特化し，A1レベルからB1レベルの日本人高校生英語学習者が用いる語彙使用の分析に焦点をあてた研究を行う必要性を感じた。

1.2 研究の背景

研究の背景として，次期学習指導要領へ向けた指導法の改善の動きを挙げることができる。その動きとは，教師中心の「指導」による学びから，学習者中心の「学習」による学びへの教育の質的転換である。次期学習指導要領では，「資質・能力」を育成するため，アクティブ・ラーニングの視点からの「主体的・対話的で深い学び」への授業改善が望まれている。特に英語教育においては，4技能を統合しながら活用できるようにするために，アクティブ・ラーニング型授業の構造化が求められている。平成28年12月に文部科学省中央教育審議会（2016）から「主体的・対話的で深い学び」についての答申が提出された。答申の本文では「協働」という語が散見されることからも分かるように，次期学習指導要領においては，学習者である児童生徒は「主体的な学び」，「対話的な学び」そして「深い学び」を実現するために，他者との「対話」や「協働（協同）」をとおしてさまざまなことを学ぶことが求められている。また，「対話」や「協働（協同）」の過程で，メタ認知（Flavell, 1979）や自己調整能力（Zimmerman, B.J., Bonner, S., & Kovach, R., 1996）を獲得することが期待されている。例えば，「読む」ことについて取り上げると，読解技能としてのモニタリングを促進する学習支援として，他の学習者との対話を通して協同で読解を行う学習法が提案されているという（佐藤，2012，p.231）。英語教師は，このような「アクティブ・ラーニング」の視点から既存の授業の見直しや再構成を図る必要がある。今，「アクティブ・ラーニング」型の学習として，脚光を浴びているのが「協働学習（特にピア・レスポンス）」で

ある。例えば，現在までにもL2ライティングとピア・インタラクションに関する研究がなされている（Storch, 2009: Philip, Adams &Iwashita, 2014:Sato and Ballinger, 2016）。こうした背景を踏まえて，本実践研究では，L2ライティング研究に，協働学習を導入するに至った。次節では，「協働学習」のうちの「ピア・レスポンス」に焦点を当てて論じる。

1.3 協働学習（ピア・レスポンス）

本実践研究における中心的な役割を果たす「協働学習（collaborative / cooperative learning）」とは，どのような学習形態のことを指すのかについて論じていきたい。「協働学習」は，厳密に言えば，他にも「協同学習（cooperative learning）」や「協調学習（collaborative learning）」（三宅・東京大学CoREF・河合塾（編），2016）と表記される場合があるが，本稿では，引用部分を除いて，便宜上すべて「協働学習」と表記することにする。

「協働学習」について，伏野（2010）では，「小グループを指導方法の１つとして使用して，学生がともに学びあいながら，自己および仲間の学びを最大限に高めることを目指す原理と技法」であると定義している。また，津田（2013, p.88）は，「数人でグループを形成して，共通の課題に取り組み，互いに意見を述べ合いながら理解を深めていく学習形態」と定義している。

Nunan（1992, p3）は，協働（collaboration）の取組によって学習者が得られる特長を次の6点にまとめている。①「学習という行為（learning）」について，学ぶことができる，あるいは，より深く学ぶことができること，②ことばに対する気づき（awareness）や，学習者自身について学ぶことができ，それ故に学習という行為について学ぶことができること，③結果として，コミュニケーションスキル（communicative skills）や，メタコミュニケーションスキル（metacommunicative skills）を開発することができること，④個人の要求や集団の要求での意見の衝突（conflicts）に直面しても，相手が満足するような言葉の表現や社交的な手続きを踏まえた表現を用いて，和解することができること，⑤学習内容と学習方法が密接に関連していることに気づくことができること，⑥意思決定に関わるタスクそれ自体を本当のコミュニケーション活動だと認識することができること，である。

以上の6点の特長を踏まえると，協働学習の本質とは，「グループで学習するが，グループ全体で答えを出せるようになればよいのではなくて，一人ひとりが学ぶ学習」であり，グループの構成員の互恵的関係が醸成されること（三宅他, 2016）であり，協同学習（協働学習）がめざす学力とは，「子どもが，主体的で自律的な学びの構え，確かで幅広い知的習得，仲間と共に課題解決に向かうことのできる対人技能，さらには，他者を尊重する民主的な態度，といった「学力」を効果的に身につけていくための「基本的な考え方」」（江利川, 2012）と指摘することができる。つまり，協働学習においては，「メンバー間の肯定的な社会的相互依存関係という特徴」（鹿毛, 2013, p.271）をいかに引き出すかが重要なのである。

以上で述べたように，協働学習においては，ヴィゴツキーの心理学の流れを汲む社会文化的アプローチの視点が取り入れられており，教室での英語学習において，ある共通の目的を持った共同体やグループによる作業，プロジェクト，作品作りなど，他者，言語，文化的道具を意識させるような活動に生徒たちが取り組むことにより，新しい参加構造を持った教室環境のデザインが可能になるという（吉田, 2001）。

「協働学習」のうち，第二言語習得研究，特にフィードバック研究によく使用されるものとしては「ピア・レスポンス」が挙げられる。「ピア・レスポンス（peer response）」とは，学習研究においては，学習者がお互いの文章を読み合って，話し合う活動（広瀬, 2007；富永, 2012）と呼ばれている。このような学習形態が，大学などの高等教育を中心に，初等中等教育においても導入，普及されるようになってきた。本実践研究においては，L2ライティング指導法であり，また学習者主体の協働学習として，ピア・レスポンスを導入した。

1.4 国内外における先行研究（ピア・レスポンスの有効性と非有効性）

協働学習の概念は，ロシアの心理学者ヴィゴツキー（Vygotsky）が提唱した最近接発達領域論

(Zone of Proximal Development：ZPD)の考えに立脚している。ZPDとは，支援なしのタスクを完遂できる学習者の能力と，より有能な相手の支援を得てタスクを完成することの間の実行できる行為の範囲のことを示している(Philip,J., et al., 2014)。

これまでにZPD理論にもとづいた協働学習(ピア・レスポンス)の有効性が研究報告されている。例えば，Ferris(2003)は，ライティングにおける教師によるレスポンスと生徒によるピア・レスポンスに関する歴史的な背景と，それらの研究の成果について論述した。また，第二言語習得に関わる言語教師や研究者にとって，L2ライティング研究は重要であるという認識を持っており，また，学習者である生徒にとっては，教師と生徒の両方からのフィードバックが最も効果的であるかもしれないという意見を主張している。Hyland, K. and Hyland, F.(2006)は，他者の行動や心的状態を予測し，観察し，解釈する能力を見る社会認知的観点から，ピア・レビュー(peer review)は，書き手が自らの文章について論じたり，他人が彼らの文章に対する解釈を行う機会を与えてくれる形成的発展過程として見なすことができると指摘している。

日本国内におけるピア・レスポンスに関する先行研究についても概観してみる。大学生を対象としたピア・レスポンス研究において，例えば，阿部・山西(2013)は，「日本人英語学習者(大学生)同士が，教師のサポートがない状況で，未知なライティングタスクに挑戦する際，協働的に学習することによってできることは何か」をヴィゴツキーの言語発達理論に依拠する社会文化的言語習得枠組みに基づき，協働学習を媒介する対話記述を，グラウンデッド・セオリー・アプローチ(Grounded Theory Approach：GTA)を用いて分析した。結果として，「学生ペアは協働的対話を通し，自らの知識を外化し，それを共有，深化させ，未知のライティングタスクに対応する知識を共に構築することができた」と報告した。また，岩中(2014)は，協同学習を取り入れた内容言語指導(Content-Language Instruction：CLI)が英語に対して苦手意識を持っている学習者(大学生)に対して有効であることを指摘した。越智(2015)は，大学生を相手にプロセス・ライティングの手法で指導したパラグラフ・ライティングをピア・エバリュエーション及びピア・レスポンス，ティーチャー・フィードバックを通して指導した結果を分析した。その結果，EFL学習者は書き手として全体の構成を意識できるようになっている可能性が認められた。Tsuchiya(2015)は，モチベーション要因低減の観点から，大学生を対象とした協同学習の効果として，協同学習が英語力の低い学生の学習に対する肯定的な態度を生み出すだけでなく，1セメスター(6か月)間，肯定的な態度を維持することに貢献したなどの有効性を挙げている。

日本人高校生を対象とした研究としては，久山(2008)，山本(2010)，野上(2016)がある。久山(2008)は，ピア・フィードバック活動を導入し，自由英作文指導におけるプロセスの改善とライティング不安の軽減の観点から指導効果の分析を行った。対象クラスの生徒を，文法能力，作文ストラテジー，英作文力の3変数によって，4つのグループに分けたが，作文ストラテジーの使用頻度が最も低かったグループにおいて，計画段階と推敲段階でのストラテジー使用が増えたと報告している。また，ライティング不安の軽減については，どのグループにおいても指導の効果は確認されなかったと報告している。山本(2010)は，日本語を使ってのピア・レスポンスがEFLライティングの質や量の改善を促すだけでなく，社会概念，人間関係，友情の育成，自己確認，文化理解など個人の社会性に影響するという結果を報告した。野上(2016)は，高校3年生を対象に，生徒同士のフィードバック活動を英作文の授業に取り入れ，構成，結束性，一貫性といった文章のまとまりに必要な要素に注目させることによって，作文能力が向上するかを検証した。その結果，英語熟達度下位群は，結束性について，実際の英作文の点数が伸びていた。しかし，フィードバックではコメントもしているにも関わらず，自信をもって使えるようになっていないことがわかった。

中学生を対象とした研究としては，神原(2011)が挙げられる。神原は，本務校の中学2年生を対象に，年間を通して6つの創作ライティングタスクを与え，教師フィードバックとともにピア・レスポンスを実施した。タスク後に実施したアンケート結果から，生徒がピア同士でアイデアを

出し合い，読み手意識を深め，内容重視のアドバイス交換ができ始めたことを報告している。また，読み手を意識したがために，複雑かつまとまりのある内容を書こうとして，文章の結束性（cohesion）に関するglobalな誤り（error）が増加したことを踏まえ，ピア・レスポンスにより，書く意欲が高まる反面，個々の生徒に応じた教師のフィードバックが重要であることを指摘した。

これまでの先行研究においては，協働学習（ピア・レスポンス）の有効性について論じられてはいたものの，実際に活動の中で創作された英文エッセイなどの作品に使用されている語彙に関して量的分析を行ったものが少なかった。特に，実際に使用されている語彙がエッセイの読み手である言語使用者のCEFRレベルにふさわしい語彙レベルであるかというプロダクト分析を行った先行研究が少ない。そのため，本実践研究においては，日本人高校生がエッセイを作成する時に使用する語彙が，読み手にふさわしいCEFRによって分類された語彙レベルであるかを分析するプロダクト研究に取り組みたいと考えた。

2 研究方法

2.1 研究の動機と目的

前任校の愛媛大学附属高等学校は，四国唯一の国立大学附属高校であり，各学年120名（3クラス編成）という小規模校である。小規模校であるがために，友達関係においても，クラス間や学年を超えて生徒同士が良好な人間関係を構築できる環境にあった。このような環境において，授業においてもピア・レスポンスを導入することで，協働学習におけるEFLライティング指導の効果を検証できるのではないかと考えた。

本実践研究では，ピア・レスポンスとそのあとに行った英文エッセイのリライト活動が，エッセイにおけるCEFRレベル別語彙使用に与える効果を分析する。具体的には協働学習，特にピア・レスポンスによる相互コメント活動後にリライトした英文エッセイプロダクトにおいて，生徒が使用する語彙にどのような質的変化や量的変化が見られるかを明らかにし，L2ライティング研究とピア・レスポンスを利用している授業活動への指針に貢献できればと思っている。本実践研究においてCEFR（Council of Europe, 2001）を分析の手段に用いた理由は，これまでの先行研究において，エッセイライティングにおける英文の質（特に語彙レベル）をCEFRに基づいて分析した先行研究が少ないことが理由である。また，CEFRのレベル付けが行われた語彙数の増減を比較することで，協働学習（ピア・レスポンス）の有効性を検証したいとも考えた。このプロダクト分析に関して，本実践研究では以下の仮説（H1）とリサーチクエスチョン（RQ1,RQ2）を重要な視点としている。仮説とリサーチクエスチョンは次のとおりである。ここで言う語彙の量的変化とは，語彙数の増減に関連することである。また，語彙の質的変化とは，CEFRに応じた語彙レベルに関する特徴の変化を指すものとする。

H1 母語を使用した相互評価活動によるフィードバックを取り入れたピア・レスポンスを行うと，生徒は相手により理解されやすいエッセイを書こうとして，リライト活動でCEFRレベルA1・A2レベルの単語をより多く使うようになる。

RQ1 協働学習，特にピア・レスポンス後のリライト活動によって，生徒が使用する語彙に量的変化が見られるか。

RQ2 協働学習，特にピア・レスポンス後のリライト活動によって，生徒が使用する語彙に質的変化が見られるか。

2.2 参加者とその学習背景

本実践研究には愛媛大学附属高等学校の1年生120名（2017年度入学：男子36名、女子84名）全員が参加した。本校は，2015年4月に文部科学省より，スーパーグローバルハイスクール（SGH）に指定されており、英語教育や国際交流，そして，愛媛大学との高大連携プログラムにも力を入れている。2年次にはSGH科目「異文化理解」を履

修することになっている。本校の海外協定校がある「アメリカ」「フィリピン」「ルーマニア」「韓国」「オーストラリア」の5か国の中から1つの国を選択することになっている。また，それらの国の「経済」「環境」「教育」「文化」「政治」の5分野から1つの分野について，自分の興味関心のあるテーマを選び，英語による課題発見・解決型の意見交換や研究交流を行う機会を持つ教育プログラムを実施している。その教育プログラムに入る前段階として，筆者は1年生に対して，コミュニケーション英語Ⅰの授業において，週1回のペースでALTとともに英語エッセイライティングを指導した。

2.3 研究活動

2.3.1 指導方法

コミュニケーション英語Ⅰでは，筆者は基本的には英語（米国式発音）を使用し，文法説明の際には英語と日本語を使用した。筆者の担当するコミュニケーション英語Ⅰ（週3コマ：1コマは45分）のうち，毎月1回（年間合計10回程度）は，ALTとのTeam Teachingの授業を組み入れた。ALTは，2016年4月に本校に採用された英国リバプール出身の40代後半男性であり，英語（英国式発音）を使って授業を行った。ALTは，本場の（authentic）イギリス英語で多量の英語を話した。生徒の聴解活動やその活動内容の理解が困難な場合にのみ，意味や形式に焦点を当てた気づき（noticing）（Schmidt,1990）を与えるために，筆者が簡単な英語で言い換え（paraphrase）を行ったり，ALTがゆっくり話したり，繰り返したりした。また，必要に応じて，筆者が発音指導を行い，実際に聞こえる発音と文字の綴りとの違いについても指導し，聴覚インプット強化（aural input enhancement）（cf. Cho, M. & Reinders, R., 2013; 長崎, 2017）を与える工夫を行った。

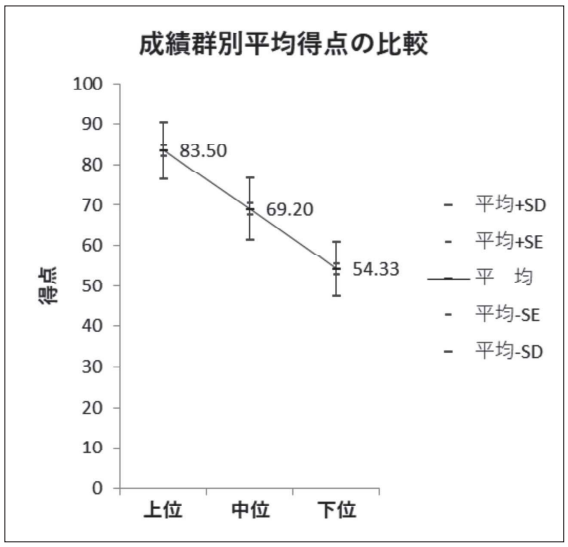
2.3.2 マテリアル

まず，参加者した生徒のうち2つのタスクを完遂した90名の対象者を各成績群（上位群30名，中位群30名，下位群30名）に分けた。3群に分けて分析を行う理由は，使用されるCEFRレベル別の語彙数や語彙の質が，学習者（高校生）の成績群に応じて，どのような違いがあるかを見るためであった。各成績群に分けるために，コミュニケーション英語Ⅰの第一学期（2017年7月実施：84問）と第二学期（2017年12月実施：65問）の期末考査試験問題を使用した。本来であれば，英検などの外部検定試験の過去問題を利用するべきところであるが，時間的に余裕がなかったため，2つの期末考査試験問題を代用した。本研究での統計分析（t検定，分散分析等）にはBellCurve for Excelを使用した。90名の生徒（対応あり）が受けた2つの試験問題の得点の相関を確認したところ，$r = .85$（$p < .001$）であり，かなり強い相関があった。また，2つの試験の得点の信頼度は$\alpha = .92$であった。この結果に基づき，第一学期期末考査の得点（素点）に基づき，参加者を上位群（30名），中位群（30名），下位群（30名）の3群に分けたところ，表1と図1のようになった。偏差は中位群が他の2群に比べて少し大きいが，標準誤差に関しては，上位群が1.27，中位群が1.37，下位群が1.23の数値であり，ほぼ同じであった。考査の平均点に関して，図1を見ると，各群の平均点を結んだ線が，ほぼ綺麗な直線を描いていることが分かる。各成績群の平均点は，上位群が83.50点，中位群が69.20点，下位群が54.33点であった。

■表1：第一学期期末考査（3群）の記述統計

Group	n	平均	標準偏差（SD）	平均-SD	平均+SD	標準誤差（SE）	平均-SE	平均+SE
上位群	30	83.50	6.98	76.52	90.48	1.27	82.23	84.77
中位群	30	69.20	7.53	61.67	76.73	1.37	67.83	70.57
下位群	30	54.33	6.73	47.60	61.67	1.23	53.10	55.56
Total	90	69.01	13.87	55.14	82.88	1.46	67.55	67.55

（注）Max=100点



■図1：成績群別平均得点の比較

これらの成績群について，それらの平均点の差が有意であるかどうかを確かめるために一元配置の分散分析（one-way ANOVA）を実施した。分散分析の適用条件として，レーベン（Levene）等分散性検定の結果を確認したところ，有意確率が，p=.8041，つまり，5％水準で有意ではないので，「3群の分散は異なる」という帰無仮説が棄却された。つまり，「3群の分散は等分散である」ことになり，分散分析の適用が可能となった。そこで，一元配置の分散分析（対応なし）を行ったところ，3群の異なり語数の平均値の差は有意であった（F（2,87）=127.662，p <.001）。その結果を表2に示した。次にどの２つの群の平均点で差が見られるかを調べるために，Tukey-Kramer法を用いた多重比較を行った。その結果，上位群と中位群（p<.001），上位群と下位群（p<.001），中位群と下位群（p<.001）の平均値にはいずれも有意な差があることがわかった。

■表2: 成績群別平均点の分散分析結果

ソース	TypeⅢ平方和	自由度	平均平方	F値	P値	**:P<.01
成績群別平均点	1276.0222	2	6381.0111	127.0662	P<.001**	
誤差	4368.9667	87	50.2180			
全体	17130.9889	89				

2.3.3 実践研究の概要

本実践研究は，コミュニケーション英語Ⅰにおいて，基本的にALTとのTeam Teachingの授業を利用して行った。研究期間は，2017年5月～2018年1月の9か月間であった。表3は研究活動計画の概要を示したものである。

1番目の活動（2017年5月第1週～第2週）では，ALTがパラグラフ・ライティングの基礎を指導した。この活動において，生徒は，A4レポート用紙（片面1枚）に1つのパラグラフから構成されるエッセイを書いた。筆者とALTとで，1つのパラグラフが，主題文（topic sentence）と複数の支持文（supporting sentences）と結論文（concluding sentence）を含めた6～7文程度で構成されるように生徒に指導した。

2番目の活動（2017年5月第3週～第5週）では，生徒は，A4レポート用紙（片面1枚）に，Introduction（導入）・Main body（本論）・Conclusion（結論）の3つパラグラフから構成されるエッセイを書いた。テーマは「私の日常生活」である。サブテーマは，

「学校生活」,「部活動」,「趣味」,「好きな食べ物」などであった。

3番目の活動(2017年6月第1週)では,2.2節で述べたように,生徒は、「アメリカ」「フィリピン」「ルーマニア」「韓国」「オーストラリア」の中から1つの国を選択した。また,その国の「経済」「環境」「教育」「文化」「政治」の分野から1つ選び,それらに関連したテーマを設定した。

4番目の活動(2017年6月第2週～第5週)では,生徒は,3番目の活動で選んだテーマに沿った調べ学習を行い,エッセイの概要を日本語で考え,その後,英語でまとめた。

5番目の活動(2017年7月)において,実際にエッセイライティングを始めた。本活動の途中で一学期の授業が終了した。エッセイの残りの部分については,夏休みの課題として提示し,英文エッセイを仕上げて,9月末に提出するように指示した。

二学期に入り,6番目の活動(2017年10月)として,ピア・レスポンスを実施した。授業5コマ分を利用し,約180分の時間をかけて行った。各自が書いた英文エッセイを5名の他の生徒と一人ずつ交換して読み合い,エッセイの構成や読みやすさについて,母語(日本語)で話し合う時間を設けた。生徒は,一人ずつ5名の他の生徒から得たコメントをワークシート(筆者が作成,資料1)に記入し

■表3: 研究活動計画

実施年月	実施週	学習内容	評価活動
2017年5月	第1週～第2週	①パラグラフ・ライティング指導	A4レポート用紙(1枚片面)に1つのパラグラフから構成されるエッセイを書く。1つのパラグラフを6～7文程度で構成する。
	第3週～第5週	②エッセイ・ライティング指導	A4レポート用紙(1枚片面)に,Introduction(導入)・Main body(本論)・Conclusion(結論)の3つのパラグラフから構成されるエッセイを書く。テーマは「私の日常生活」
2017年6月	第1週	③国・分野別テーマ選択	アメリカ,フィリピン,ルーマニア,韓国,オーストラリアの中から1つの国を選択。またその国の「経済」「環境」「教育」「文化」「政治」の分野から1つ選び,関連したテーマを設定する。
2017年6月	第2週～第5週	④テーマ別調べ学習	テーマに沿った調べ学習を行い,エッセイの概要を考える。
2017年7月	第1週～第5週	⑤プレタスク(エッセイ・ライティング)	A4用紙(片面)英文エッセイ(総語数250語程度)夏休み中は自由に好きな時間だけ取り組む。
2017年8月	第1週～第5週		
2017年9月	第1週～第5週		
2017年10月	第1週～第5週	⑥ピア・フィードバック	ペアによるフィードバック。毎週1人×5週=5人からフィードバックをもらう。
2017年11月	第1週～第5週	⑦ポストタスク(リライト活動)	A4用紙(片面)英文エッセイ(総語数350語～400語程度)夏休み課題のエッセイをリライトする。
2017年12月			
2018年1月			

た。生徒は相手の生徒のコメントを読み直し，自分のエッセイの主題文の有無であるとか，それぞれのパラグラフがIntroduction, Main body, Conclusionのいずれの部分に相当するのかということを考えるなど，生徒同士によるフィードバックからなるべく多くの気づきを得るための活動を行った。

7番目の活動（2017年11月～2018年1月）では，生徒たちは，他の5名の生徒から得たフィードバックに基づいて，エッセイのリライト活動を行った。この時，筆者やALTは生徒の学習活動への介入を避けるため，生徒への明示的な誤り訂正（explicit error correction）は行わなかった。プレタスクでは英文エッセイが250語程度を，ポストタスクでは350語以上400語程度の長さの英文エッセイを書くように指示しただけであった。生徒は，自分自身で考えながら，ポストタスクの課題エッセイに取り組み，冬休み明けに完成させた。生徒はプレタスクとポストタスクのいずれの英文エッセイも，マイクロソフト社のWordで作成した電子データをメールで筆者に送付するとともに，紙媒体でも提出した。

2.3.4 データ分析方法

筆者は，エッセイ中の語彙をCEFRレベル別に分類するために，インターネットで公開されている英国English ProfileのText Inspector（参考URL：http://www.englishprofile.org/wordlists/text-inspector）を使用した。English Profileとは，協働的（collaborative）で，研究，参照，出版の学際的なプログラムであり，CEFR（ヨーロッパ言語共通参照枠）が，言語の学習，教授，そして評価に関して何を意味しているのかを詳細に説明するために設計されたものである（Harrison, 2015）。また，Text Inspectorは，CEFRの英語への導入を研究しているEnglish Profileによって公開されており，「英文テキストの語彙のCEFRレベルと出現率を縦棒グラフで表示することができる分析プログラム」（三好, 2016）であり，500語まで無料で分析することができる。

本実践研究では，参加者（120名）のうち，夏休み課題エッセイのプレタスク（以下，プレタスク）とプレタスクをリライトさせた冬休み課題エッセイのポストタスク（以下，ポストタスク）の2つのタスクを完遂した日本人高校生英語学習者（90名：男子24名，女子66名）のCEFRレベル別語彙使用数をText Inspectorによって分析した。最終的にプレタスクとポストタスクで得られたデータを，分散分析と多重比較によって，ピア・レスポンスによる学習効果を検証した。

まず，夏休み後に提出させたプレタスクの英文エッセイの記述データを処理し，全体の傾向を見た。表4は，プレタスクの異なり語数（Pre-Task_Types）の記述統計量であり，表5はプレタスクの延べ語数（Pre-Task_Tokens）に関する記述統計量を示している。天井効果またはフロア効果については網かけしてある。表4において，プレタスクの異なり語数の総語数（Pre-types_total_SUM）は，13,037語（n=90）で，一人あたり平均（Pre-types_total_M），144.85語の異なり語数を使用して書いていた。そのうち，A1レベルの異なり語数の平均値（Pre-A1_types_M）は，80.03語，A2レベルの異なり語数の平均値（Pre-A2_types_M）は，23.68語であり，CEFR　Aレベル（A1・A2）のみの語彙に着目すると，全体の語彙の71.6％を占めていることが分かった。

表5において，プレタスクの延べ総語数（Pre-Task_Tokens）の平均値（Pre-tokens_total_M）は309.25語で，各生徒が310語程度のエッセイを書いたことが分かる。プレタスクで指示した250語を大きく上回っていた。このプレタスクの事前調査に関して，表4のC2レベルの異なり語数の偏差（C2_types_M-SD）では，かなりのフロア効果を示している。また，表5のC2レベルの延べ語数の偏差（C2_tokens_M-SD）でも同様の傾向が見られた。このことから，C2レベルの語彙は，全体的な偏りから言うと，高校生1年生の段階ではほぼ書けていない，あるいは書けないということになる。表4において，A1レベルの異なり語の総語数（Pre-A1_types_SUM）と次に多いA2レベルの異なり語数（Pre-A2_types_SUM）を比較したところ，A1レベルのtypes_SUMの方がA2レベルのtypes_SUM よりも3倍近く多いことが分かった。

■表4: プレタスクの異なり語数(types)の記述統計量

	Pre-Task_Types（n=90）						
	SUM	*M*	*SD*	*M+SD*	*M-SD*	min.	max.
A1_types	7203	80.03	11.16	91.19	68.87	45	104
A2_types	2131	23.68	5.63	29.31	18.05	9	38
B1_types	1489	16.54	6.43	22.97	10.11	3	40
B2_types	848	9.42	6.02	15.44	3.40	1	34
C1_types	217	2.41	1.62	4.03	0.79	0	8
C2_types	88	0.98	1.13	2.11	-0.15	0	5
Unlisted	1061	11.79	7.37	19.16	4.42	2	42
total	13037	144.85	18.07	162.92	126.78	111	209

■表5: プレタスクの延べ語数(tokens)の記述統計量

	Pre-Task_Tokens（n=90）						
	SUM	*M*	*SD*	*M+SD*	*M-SD*	min.	max.
A1_tokens	19290	214.33	28.94	243.27	185.39	138	293
A2_tokens	3147	34.97	10.77	45.74	24.20	16	89
B1_tokens	2039	22.66	9.07	31.73	13.59	4	54
B2_tokens	1124	12.49	8.73	21.22	3.76	1	41
C1_tokens	351	3.90	3.52	7.42	0.38	0	18
C2_tokens	99	1.10	1.34	2.44	-0.24	0	6
Unlisted	1782	19.80	12.83	32.63	6.97	0	62
total	27832	309.25	30.45	339.70	278.80	228	431

3 分析結果

3.1 CEFRレベル別異なり語数(types)について

本節では，CEFRレベル別の異なり語数について分析していく。表6は，ポストタスクにおける各CEFRレベル別の異なり語数(Post-Task_Types)の記述統計量であり，表7はポストタスクにおける各CEFRレベル別の延べ語数(Post-Task_Tokens)に関する記述統計量を示している。天井効果またはフロア効果については網かけしてある。表6において，ポストタスクの異なり語数の総語数(Post-types_total_SUM)は，14,932語(n=90)で，一人あたりの異なり語数の平均値(Post-types_total_M)は165.91語であることが分かった。

ここで，表4のプレタスクの各CEFRレベル別の異なり語数の総語数(Pre-types_total_SUM)と表6のポストタスクの同語数(Post-types_total_SUM)との比較を行った。プレタスクの異なり語数の総語数(Pre-types_total_SUM)13,037語(n=90)との比較では全体的に1,895語の増加による差があった。また，プレタスクの一人あたりの異なり語数の平均値(Pre-types_total_M)である144.85語と比較すると，ポストタスクの一人あたりの異なり語数の平均値(Post-types_total_M)との差は，21.06語の増加であった。

■表6: ポストタスクの異なり語数(types)の記述統計量

	Post-Task_Types（n=90）						
	SUM	*M*	*SD*	*M+SD*	*M-SD*	min.	max.
A1_types	8026	89.18	14.38	103.56	74.80	54	119
A2_types	2563	28.48	5.89	34.37	22.59	13	43
B1_types	1800	20.00	7.79	27.79	12.21	7	46
B2_types	984	10.93	7.42	18.35	3.51	2	40
C1_types	244	2.71	2.06	4.77	0.65	0	9
C2_types	98	1.09	1.40	2.49	-0.31	0	6
Unlisted	1217	13.52	8.05	21.57	5.47	2	51
total	14932	165.91	23.45	189.36	142.46	112	230

表7において，ポストタスクの延べ総語数の平均値（Post-tokens_total_M）は374.18語で，各生徒が375語程度のエッセイを書いたことになり，プレタスクと比較すると，65語程度増加していた。

このポストタスクに関して，表6のC2レベルの異なり語数の偏差（C2_types_M-SD）が，プレタスクの同語数の偏差よりも，かなりのフロア効果が見られた。また，表7のC2レベルの延べ語数の偏差（C2_tokens_M-SD）でも同様の傾向が見られた。プレタスクとポストタスクC2レベルの異なり語数の偏差（C2_types_M-SD）に着目し，比較すると，ポストタスクの方がより，偏差がよりマイナスの値をとっていることから，リライト活動によって，C2レベルの語彙の使用が減少しているということが言えるだろう。

■表7: ポストタスクの延べ語数(tokens)の記述統計量

	Post-Task_Tokens（n=90）						
	SUM	*M*	*SD*	*M+SD*	*M-SD*	min.	max.
A1_tokens	23308	258.98	42.98	301.96	216.00	107	337
A2_tokens	3899	43.32	10.67	53.99	32.65	19	73
B1_tokens	2519	27.99	11.02	39.01	16.97	8	63
B2_tokens	1334	14.82	10.13	24.95	4.69	2	50
C1_tokens	390	4.33	4.17	8.50	0.16	0	17
C2_tokens	104	1.16	1.58	2.74	-0.42	0	7
Unlisted	2122	23.58	14.14	37.72	9.44	3	63
total	33676	374.18	47.28	421.46	326.90	213	476

再度，表4のプレタスクの異なり語数の総語数（Pre-types_total_SUM）と表6のポストタスクの異なり語数の総語数（Post-types_total_SUM）を比較したとき，全体的に見れば，生徒が書いたプロダクトの英文量は増加していると言えそうであるが，プレタスクの異なり総語数とポストタスクの異なり総語数の差が統計的に有意かどうかを確かめるために，有意水準 5% で両側検定のt検定を行った。t 検定の結果，t(90)= 10.057, df=89, p<.001, 効果量d= 1.0059であり，プレタスクとポストタスクの異なり総語数の差は有意であることがわかった（表8）。したがって，異な

■表8: 対応のあるt検定の結果(異なり総語数の差)

	異なり総語数の(対応あり)の差 (n=90)							
	平均値 Mean	標準偏差 SD	効果量 *d*	差の平均の95%信頼区間		t値	自由度 *df*	有意確率(両側) *p*
				下限	上限			
プレタスクとポストタスクの差	21.05	19.86	1.0059	16.8956	25.2155	10.057	89	<.001

り語数に関して, プレタスクからポストタスクにかけて増加したことから, 生徒のライティングによるアウトプット量が増えただけでなく, 多様な語彙を用いてエッセイを書いたという点で, 語彙の豊かさも増したと言えるだろう。

次に, プレタスクとポストタスクにおけるプロダクトのCEFR A1レベルの異なり総語数の平均値の差についても同様に分析を行った。表6のポストタスクのA1レベルの異なり総語数の平均値(Post-A1_types_M)は, 89.18語, A2レベルの異なり総語数の平均値(Post-A2_types_M)は, 28.48語であり, CEFR Aレベル(A1・A2)のみの語彙に着目すると, 全体の語彙の70.9%を占めていることが分かった。そこで, 表4のプレタスクのCEFR A1レベルの異なり総語数の平均値(Pre-A1_types_M)とA2レベルの異なり総語数の平均値(Pre-A2_types_M)とを, 表6のポストタスクのA1とA2レベルの同語数の平均値と比較すると, ポストタスクのA1レベルでは, 9.14語の増加, A2レベルでは, 4.80語の増加が見られた。プレタスクとポストタスクのCEFR A1レベル同士の異なり総語数の平均値の差, そしてCEFR A2レベル同士の異なり総語数の平均値の差が統計的に有意かどうかを確かめるために, 有意水準5%で両側検定のt検定を行った。

t検定の結果, プレタスクとポストタスクのCEFR A1レベル同士の異なり総語数の平均値の差は, t(90)= 9.227, df=89, p<.001, 効果量d= .7104であった。また, CEFR A2レベルの異なり総語数の差は, t(90)=9.175, df=89, p<.001, 効果量d= .8333であった。以上の結果から, CEFR Aレベル(A1・A2ともに)の異なり総語数の平均値の差は有意であることが確認された(表9)。

■表9: 対応のあるt検定の結果(異なり総語数[CEFR A1・A2 レベル]の平均値の差)

	CEFR A レベル総語数(対応あり)の平均値の差 (n=90)							
	平均値 Mean	標準偏差 SD	効果量 *d*	差の平均の95%信頼区間		t値	自由度 *df*	有意確率(両側) *p*
				下限	上限			
プレタスクとポストタスクの差(A1レベル)	9.14	9.40	0.7104	7.1753	11.1136	9.227	89	<.001
プレタスクとポストタスクの差(A2レベル)	4.80	4.96	0.8333	3.7605	5.8395	9.175	89	<.001

プレタスクとポストタスクのCEFR Bレベルの異なり語数の総語数の平均値の差についても, 同様の検証を行った。表4のプレタスクのB1レベルの異なり語数の総語数の平均値(Pre-B1_types_M)は, 16.54語であり, 表6のポストタスクのB1レベルの同語数の平均値(Post-B1_types_M)は, 20.00語であった。両者の平均値を比較すると, 3.46語の増加であった。また, プレタスクのB2レベルの異なり語数の総語数の平均値(Pre-B2_types_M)は9.42語であり, ポストタスクのB2レベルの同語数の平均値(Post-B2_types_M)は, 10.93語であった。両者の平均値を

比較すると，1.51語の増加であった。プレタスクとポストタスクのCEFR　B1レベル同士の異なり総語数の平均値の差，そしてCEFR　B2レベル同士の異なり総語数の平均値の差が統計的に有意かどうかを確かめるために，有意水準 5%で両側検定のt検定を行った。

t 検定の結果，プレタスクとポストタスクのCEFR　B1レベル同士の異なり総語数の平均値の差は，t(90)= 6.790, df=89, p<.001, 効果量d= .4837であった。また，CEFR B2レベルの異なり総語数の平均値の差は，t(90)=3.462、df=89、p<.001, 効果量d= .2237であった。以上の結果から，CEFR B1レベル同士の総語数の平均値の差は，有意である言えるが，CEFR B2レベル同士の総語数の平均値の差は，効果量の大きさから考えると有意ではあるが，その差は小さいということが言える(表10)。

■表10: 対応のあるt検定の結果(異なり総語数[CEFR B1・B2 レベル]の平均値の差)

	CEFR B レベル総語数(対応あり)の平均値の差 (n=90)							
	平均値 Mean	標準偏差 SD	効果量 *d*	差の平均の95%信頼区間		t値	自由度 *df*	有意確率(両側) *p*
				下限	上限			
プレタスクとポストタスクの差(B1レベル)	3.46	4.82	0.4837	2.4444	4.4667	6.790	89	<.001
プレタスクとポストタスクの差(B2レベル)	1.51	4.14	0.2237	0.6437	2.3785	3.462	89	<.001

プレタスクとポストタスクのCEFR Cレベルの異なり語数の総語数の平均値の差についても，同様の検証を行った。表4のプレタスクのC1レベルの異なり語数の総語数の平均値(Pre-C1_types_M)は，2.41語であり，表6のポストタスクのC1レベルの同語数の平均値(Post-C1_types_M)は，2.71語であった。両者の平均値を比較すると，0.30語の増加であった。また，プレタスクのC2レベルの異なり語数の総語数の平均値(Pre-C2_types_M)は0.98語であり，ポストタスクのC2レベルの同語数の平均値(Post-C2_types_M)は，1.09語であった。両者の平均値の差を比較すると，0.11語の増加であったが，これらは，ほとんど差がないように思われた。そこで，プレタスクとポストタスクのCEFR C1レベル同士の異なり総語数の平均値の差，そしてCEFR　C2レベル同士の異なり総語数の平均値の差が統計的に有意かどうかを確かめるために，有意水準 5%で両側検定のt検定を行った。

t 検定の結果，プレタスクとポストタスクのCEFR　C1レベル同士の異なり総語数の平均

■表11: 対応のあるt検定の結果(異なり総語数[CEFR C1・C2レベル]の平均値の差)

	CEFR C レベル総語数(対応あり)の平均値の差 (n=90)							
	平均値 Mean	標準偏差 SD	効果量 *d*	差の平均の95%信頼区間		t値	自由度 *df*	有意確率(両側) *p*
				下限	上限			
プレタスクとポストタスクの差(C1レベル)	0.30	1.67	0.1621	-0.0488	0.6488	1.7087	89	.091
プレタスクとポストタスクの差(C2レベル)	0.11	0.87	0.0877	-0.0705	0.2927	1.2159	89	.227

値の差は，t（90） = 1.7087， df=89, p=.091，効果量d= .1621であった。また、CEFR C2レベルの異なり総語数の差は，t（90）=1.2159，df=89, p=.227，効果量d= .0877であった。以上の結果から，CEFR C1レベル同士とC2レベル同士の総語数の平均値の差は，有意でないことがわかった（表11）。つまり，CEFR C1レベルやC2レベルの語彙数は，リライト前後であまり変化することがないことを示している。

3.2 CEFRレベル別延べ語数（tokens）について

本節では，CEFRレベル別延べ語数について分析した結果を述べる。プレタスクとポストタスクにおけるプロダクトのCEFR　A1からC2までの各レベルの延べ総語数の平均値の差について分析を行った。表5より，プレタスクのA1レベルの延べ総語数の平均値（Pre-A1_tokens_M）は，214.33語，A2レベルの延べ総語数の平均値（Pre-A2_tokens_M）は，34.97語であり，CEFR　Aレベル（A1・A2）のみの語彙に着目すると，全体の語彙の80.6%を占めていることが分かった。そこで，表7のポストタスクのCEFR A1レベルの延べ総語数の平均値（Post-A1_tokens_M）とA2レベルの延べ総語数の平均値（Post-A2_tokens_M）とを，それぞれ表4のプレタスクのA1とA2レベルの同語数の平均値と比較すると，ポストタスクのA1レベルでは，44.65語の増加，A2レベルでは，8.35語の増加が見られた。プレタスクとポストタスクのCEFR　A1レベル同士の延べ総語数の平均値の差，そしてCEFR　A2レベル同士の延べ総語数の平均値の差がそれぞれ統計的に有意かどうかを確かめるために，有意水準5%で両側検定のt検定を行った。t検定の結果，プレタスクとポストタスクのCEFR　A１レベル同士の延べ総語数の平均値の差は，t（90）= 11.9506，df=89, p<.001，効果量d= 1.2185であった。また，CEFR A2レベルの延べ総語数の差は，t（90）=8.9122，df=89, p<.001，効果量d= .7791であった。以上の結果から，CEFR Aレベル（A1・A2）の延べ総語数の平均値の差は有意であることが確認された（表12）。

同様にプレタスクとポストタスクのCEFR Bレベルの延べ総語数の平均値の差についても，同様の検証を行った。表5のプレタスクのB1レベルの延べ総語数の平均値（Pre-B1_tokens_M）は，22.66語であり，表7のポストタスクのB1レベルの同語数の平均値（Post-B1_tokens_M）は，27.99語であった。これらの平均値を比較すると，5.33語の増加であった。表5のプレタスクのB2レベルの延べ総語数の平均値（Pre-B2_tokens_M）は12.49語であり，表7のポストタスクのB2レベルの同語数の平均値（Post-B2_tokens_M）は，14.82語であった。両者の平均値を比較すると，2.33語の増加であった。プレタスクとポストタスクのCEFR　B1レベル同士の延べ総語数の平均値の差，そしてCEFR　B2レベル同士の延べ総語数の平均値の差が統計的に有意かどうかを確かめるために，有意水準5%で両側検定のt検定を行った。t検定の結果，プレタスクとポストタスクのCEFR　B1レベル同士の延べ総語数の平均値の差は，t（90）= 8.0316，df=89, p<.001，効果量d= .5293であった。また，CEFR B2レベルの延べ総語数の差は，t（90）=3.9550，df=89, p<.001，効果量d= .2468であった。以上の結果から，CEFR B1レベル同士の総語数の平均値の差は，有意である言えるが，CEFR B2レベル同士の総語数の平均値の差は，効果量の大きさから考えると有意ではあるが，その差は小さいということが言える（表12）。

また，プレタスクとポストタスクのCEFR Cレベルの延べ総語数の平均値の差についても，同様の検証を行った。表5のプレタスクのC1レベルの延べ総語数の平均値（Pre-C1_tokens_M）は，3.90語であり，表7のポストタスクのC1レベルの同語数の平均値（Post-C1_tokens_M）は，4.33語であった。これらの平均値を比較すると，0.43語の増加であった。この結果から，C1レベルではプレタスクとポストタスクの延べ総語数の平均値を比べた場合，ほぼ差がないように思われる。また，表5のプレタスクのC2レベルの延べ総語数の平均値（Pre-C2_tokens_M）は1.10語であった。表7のポストタスクのC2レベルの同語数の平均値（Post-C2_tokens_M）は，1.16語であった。両者の平均値の差を見ると，0.06語の増加であり，C1同様にほぼ差がないように見える。プレタスクとポストタスクのCEFR　C1

レベル同士の延べ総語数の平均値の差，そしてCEFR　C2レベル同士の延べ総語数の平均値の差が統計的に有意かどうかを確かめるために，有意水準 5%で両側検定のt検定を行った。t 検定の結果，プレタスクとポストタスクのCEFR C 1 レベル同士の延べ総語数の平均値の差は，t (90) = 1.1476，df=89, p=.245，効果量d= .1124であった。また，CEFR C2レベルの延べ総語数の平均値の差については，t(90) =.5534，df=89, p=.581，効果量d= .0379であった。以上の結果から，CEFR C1レベル同士，そしてC2レベル同士の延べ総語数の平均値の差は，有意であるとは言えず，CEFR C1レベルとC2レベルの延べ総語数の平均値の差はないことがわかった。つまり，CEFR C1レベルやC2レベルの延べ総語数は，リライト前後であまり変化することがないことを示している(表12)。

3.3 成績群別CEFR A1レベル異なり語数(types)の平均値の差について

本節では，成績群別(上位群・中位群・下位群)に見たときに，CEFRレベル別の異なり総語数の平均値について差が見られるかどうかを検証することにした。まず，CEFR　A1レベルの異なり総語数の平均値の差について，調査を実施した(表13・図2)。分散分析の適用条件として，レーベン(Levene)等分散性検定の結果を確認したところ，有意確率が，p=.5057，つまり，5%水準で有意ではないので，「3群の分散は異なる」という帰無仮説が棄却された。つまり，「3群の分散は等分散である」ことになり，分散分析の適用が可能となった。そこで，一元配置の分散分析を行ったところ，3群の異なり語数の平均値の差は有意であった(F (5,174)) =4.7205，p <.001)。その結果を表14に示した。次にどの2つの群の平均値の間で差が見られるかを調べるために，Tukey-Kramer法を用いた多重比較を行った。その結果，プレタスクに取り組んだ成績中位群の生徒の異なり語数の平均値とポストタスクに取り組んだ成績上位群の生徒の異なり語数の平均値には有意差があることがわかった(p<.05)。また，プレタスクの下位群とポストタスクの上位群にも有意差が見られた(p<.01)。しかしながら，CEFR　A1レベルでは，上位群，中位群，下位群の各群については，プレタスクとポストタスクで平均値の差に有意性は見られなかった(表15)。

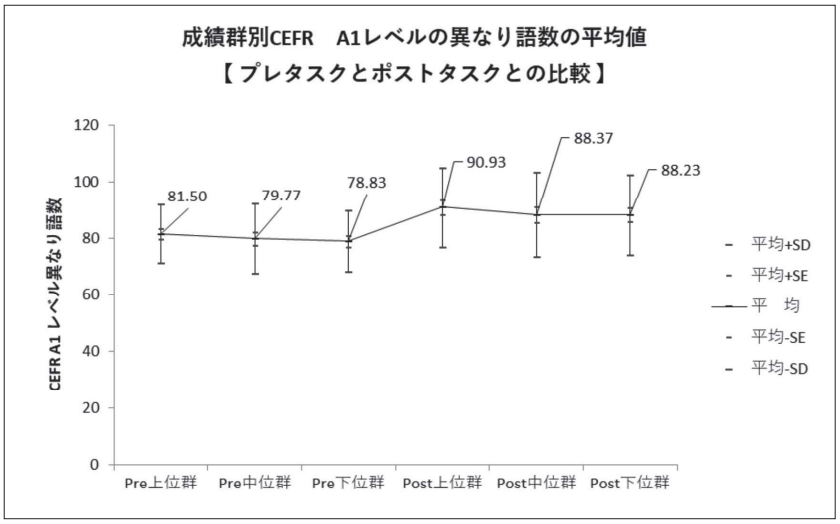
■表12: 対応のあるt検定の結果(CEFRレベル別延べ総語数の平均値の差)

	CEFR レベル別延べ総語数(対応あり)の平均値の差 (n=90)							
	平均値 Mean	標準偏差 SD	効果量 *d*	差の平均の95%信頼区間		t値	自由度 *df*	有意確率(両側) *p*
				下限	上限			
プレタスクとポストタスクの差(A1レベル)	44.65	35.44	1.2185	37.2216	52.0673	11.9506	89	**<.001
プレタスクとポストタスクの差(A2レベル)	8.35	8.89	0.7791	6.4927	10.2184	8.9122	89	**<.001
プレタスクとポストタスクの差(B1レベル)	5.33	6.30	0.5293	4.0139	6.6528	8.0316	89	**<.001
プレタスクとポストタスクの差(B2レベル)	2.33	5.60	0.2468	1.1611	3.5056	3.9550	89	**<.001
プレタスクとポストタスクの差(C1レベル)	0.43	3.58	0.1124	-0.3170	1.1836	1.1476	89	.254
プレタスクとポストタスクの差(C2レベル)	0.06	0.95	0.0379	-0.1439	0.2550	0.5534	89	.581

Note: **:P<.01

■表13: 成績群別CEFR A1レベルの異なり語数(平均値)

成績群	n	平均	標準偏差(SD)	平均-SD	平均+SD	標準誤差(SE)	平均-SE	平均+SE
Pre上位群	30	81.50	10.43	71.07	91.93	1.90	79.60	83.40
Pre中位群	30	79.77	12.46	67.31	92.22	2.27	77.49	82.04
Pre下位群	30	78.83	10.90	67.93	89.74	1.99	76.84	80.82
Post上位群	30	90.93	14.23	76.70	105.17	2.60	88.33	93.53
Post中位群	30	88.37	15.19	73.18	103.55	2.77	85.59	91.14
Post下位群	30	88.23	14.28	73.95	102.51	2.61	85.63	90.84



■図2:成績群別CEFR A1レベルの異なり語数の平均値比較

■表14: 分散分析結果

ソース	TypeⅢ平方和	自由度	平均平方	F値	P値	**:P<.01
成績群	4011.761	5	802.352	4.7205	P<.001**	
誤差	29575.233	174	169.973			
全体	33586.994	179				

■表15: 多重比較の結果(CEFR A1 レベルの異なり語数)

	Pre上位群	Pre中位群	Pre下位群	Post上位群	Post中位群	Post下位群
Pre上位群	–	.9955	.9685	.0618	.3243	.3464
Pre中位群		–	.9998	.0139*	.1143	.1253
Pre下位群			–	.0056**	.0571	.0634
Post上位群				–	.9734	.9668
Post中位群					–	1.0000
Post下位群						–

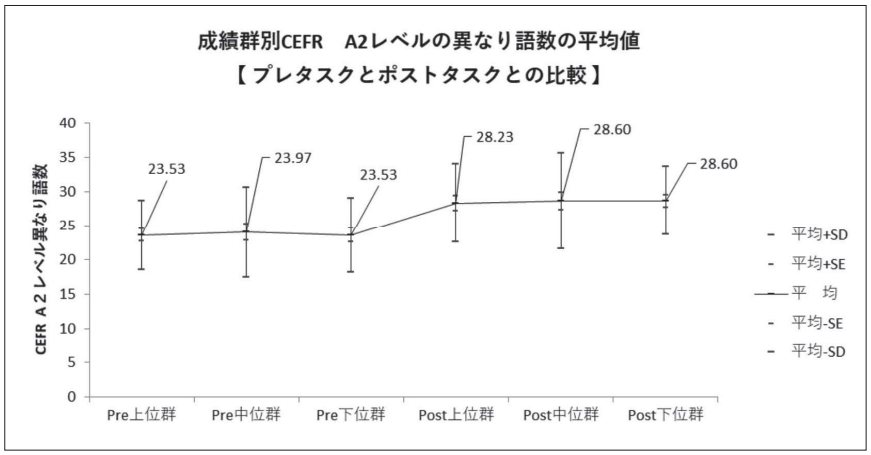
Note: *P<.05, **P<.01

3.4 成績群別CEFR A2レベル異なり語数(types)の平均値の差について

本節では，CEFR A2レベルの異なり総語数の平均値の差について，調査を実施した(表16・図3)。分散分析の適用条件として，レーベン(Levene)等分散性検定の結果を確認したところ，有意確率が，p=.5049，つまり，5%水準で有意ではないので，「3群の分散は異なる」という帰無仮説が棄却された。つまり，「3群の分散は等分散である」ことになり，分散分析の適用が可能となった。そこで，一元配置の分散分析を行ったところ，3群の異なり語数の平均値の差は有意であった(F(5,174))=6.0856，p <.001)。その結果は表17である。次にTukey-Kramer法を用いた多重比較を行った結果，表18よりプレタスクに取り組んだ成績上位群の生徒の異なり語数の平均値とポストタスクに取り組んだ成績上位群の生徒の異なり語数の平均値には有意差があることがわかった(p<.05)。また，プレタスクの中位群とポストタスクの中位群にも有意差が見られた(p<.05)。プレタスクの下位群とポストタスクの下位群にも有意差があった(p<.05)。以上から，上位群，中位群，下位群の各群について，プレタスクとポストタスクで有意差が見られた。

■表16: 成績群別CEFR A2レベルの異なり語数(平均値)

成績群	n	平均	標準偏差(SD)	平均-SD	平均+SD	標準誤差(SE)	平均-SE	平均+SE
Pre上位群	30	23.53	5.03	18.50	28.56	0.92	22.62	24.45
Pre中位群	30	23.97	6.58	17.39	30.55	1.20	22.77	25.17
Pre下位群	30	23.53	5.44	18.09	28.98	0.99	22.54	24.53
Post上位群	30	28.23	5.75	22.49	33.98	1.05	27.18	29.28
Post中位群	30	28.60	7.01	21.59	35.61	1.28	27.32	29.88
Post下位群	30	28.60	5.04	23.56	33.64	0.92	27.68	29.52



■図3:成績群別CEFR A2レベルの異なり語数の平均値比較

■表17: 分散分析結果

ソース	TypeⅢ平方和	自由度	平均平方	F値	P値	**:P<.01
成績群	1043.244	5	208.649	6.0856	P<.001**	
誤差	5965.667	174	34.285			
全体	7008.911	179				

■表18: 多重比較の結果(CEFR A2 レベルの異なり語数)

	Pre上位群	Pre中位群	Pre下位群	Post上位群	Post中位群	Post下位群
Pre上位群	–	.9997	1.0000	.0263*	.0125*	.0125*
Pre中位群		–	.9997	.0587	.0299*	0.299*
Pre下位群			–	.0263*	.0125*	.0125*
Post上位群				–	.9999	.9999
Post中位群					–	1.0000
Post下位群						–

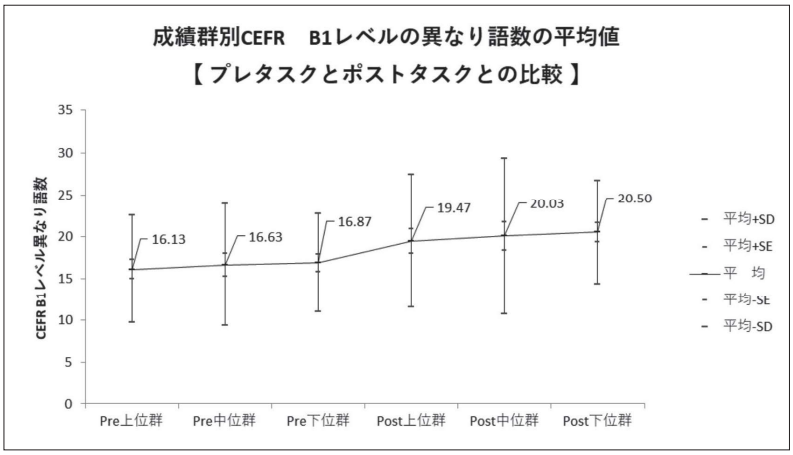
Note: *P<.05, **P<.01

3.5 成績群別CEFR B1レベル異なり語数(types)の平均値の差について

次にCEFR　B1レベルの異なり総語数の平均値の差について, 調査を実施した(表19・図4)。レーベン(Levene)等分散性検定の結果を確認したところ, 有意確率が, p=.3381, つまり, 5%水準で有意ではないので,「3群の分散は異なる」という帰無仮説が棄却された。つまり,「3群の分散は等分散である」ことになり, 分散分析の適用が可能となった。そこで, 一元配置の分散分析を行ったところ, 3群の異なり語数の平均値の差に有意性はなかった(F(5,174))=2.1340, n.s.)。その結果は表20である。次にTukey-Kramer法を用いた多重比較を行った結果, CEFR　B1レベルではプレタスクとポストタスクともに成績群に関係なく異なり語数の平均値に有意な差は見られなかった(表21)。

■表19: 成績群別CEFR B1レベルの異なり語数(平均値)

成績群	n	平均	標準偏差(SD)	平均-SD	平均+SD	標準誤差(SE)	平均-SE	平均+SE
Pre上位群	30	16.13	6.39	9.74	22.53	1.17	14.97	17.30
Pre中位群	30	16.63	7.27	9.36	23.90	1.33	15.31	17.96
Pre下位群	30	16.87	5.85	11.01	22.72	1.07	15.80	17.94
Post上位群	30	19.47	7.94	11.52	27.41	1.45	18.02	20.92
Post中位群	30	20.03	9.32	10.71	29.35	1.70	18.33	21.73
Post下位群	30	20.50	6.16	14.34	26.66	1.13	19.37	21.63



■図4:成績群別CEFR B1レベルの異なり語数の平均値比較

■表20: 分散分析結果

ソース	TypeⅢ平方和	自由度	平均平方	F値	P値
成績群	561.828	5	112.366	2.1340	0.0636
誤差	9161.833	174	52.654		
全体	9723.661	179			

■表21: 多重比較の結果(CEFR B1レベルの異なり語数)

	Pre上位群	Pre中位群	Pre下位群	Post上位群	Post中位群	Post下位群
Pre上位群	–	.9998	.9988	.4819	.3018	.1875
Pre中位群		–	1.0000	.6569	.4590	.3113
Pre下位群			–	.7344	.5401	.3819
Post上位群				–	.9997	.9939
Post中位群					–	.9999
Post下位群						–

Note: *P<.05, **P<.01

3.6 成績群別CEFR B2レベル異なり語数(types)の平均値の差について

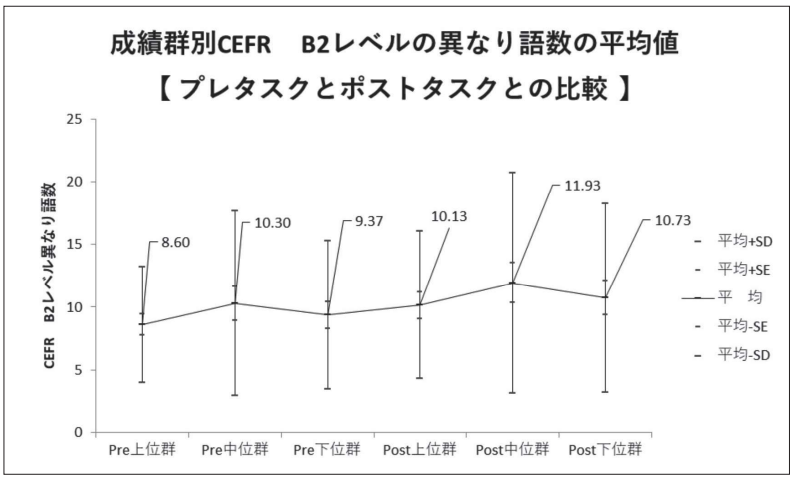
CEFR　B2レベルの異なり総語数の平均値の差について，調査を実施した(表22・図5)。分散分析の適用条件として，レーベン(Levene)等分散性検定の結果を確認したところ，有意確率が，p=.1200，つまり，5%水準で有意ではないので，「3群の分散は異なる」という帰無仮説が棄却された。つまり，「3群の分散は等分散である」ことになり，分散分析の適用が可能となった。次に一元配置の分散分析を行ったところ，3群の異なり語数の平均値の差は有意ではない(F(5,174))=.8428, n.s.)。つまり，差はないことが分かった。その結果は表23である。Tukey-Kramer法を用いた多重比較を行った結果においても，ポストタスクとプレタスクで，成績群別に見ても，有意な差は見られなかった(表24)。

■表22: 成績群別CFFR B2レベルの異なり語数(平均値)

成績群	n	平均	標準偏差(SD)	平均-SD	平均+SD	標準誤差(SE)	平均-SE	平均+SE
Pre上位群	30	8.60	4.63	3.97	13.23	0.84	7.76	9.44
Pre中位群	30	10.30	7.39	2.91	17.69	1.35	8.95	11.65
Pre下位群	30	9.37	5.92	3.45	15.28	1.08	8.29	10.45
Post上位群	30	10.13	5.89	4.24	16.03	1.08	9.06	11.21
Post中位群	30	11.93	8.80	3.13	20.74	1.61	10.33	13.54
Post下位群	30	10.73	7.53	3.20	18.27	1.38	9.36	12.11

■表23: 分散分析結果

ソース	TypeⅢ平方和	自由度	平均平方	F値	P値
成績群	196.644	5	39.329	0.8428	0.5211
誤差	8119.667	174	46.665		
全体	8316.311	179			



■図5：成績群別CEFR B2レベルの異なり語数の平均値比較

■表24: 多重比較の結果(CEFR B2レベルの異なり語数)

	Pre上位群	Pre中位群	Pre下位群	Post上位群	Post中位群	Post下位群
Pre上位群	–	.9287	.9980	.9533	.4119	.8318
Pre中位群		–	.9949	1.0000	.9393	.9999
Pre下位群			–	.9980	.6931	.9714
Post上位群				–	.9106	.9994
Post中位群					–	.9839
Post下位群						–

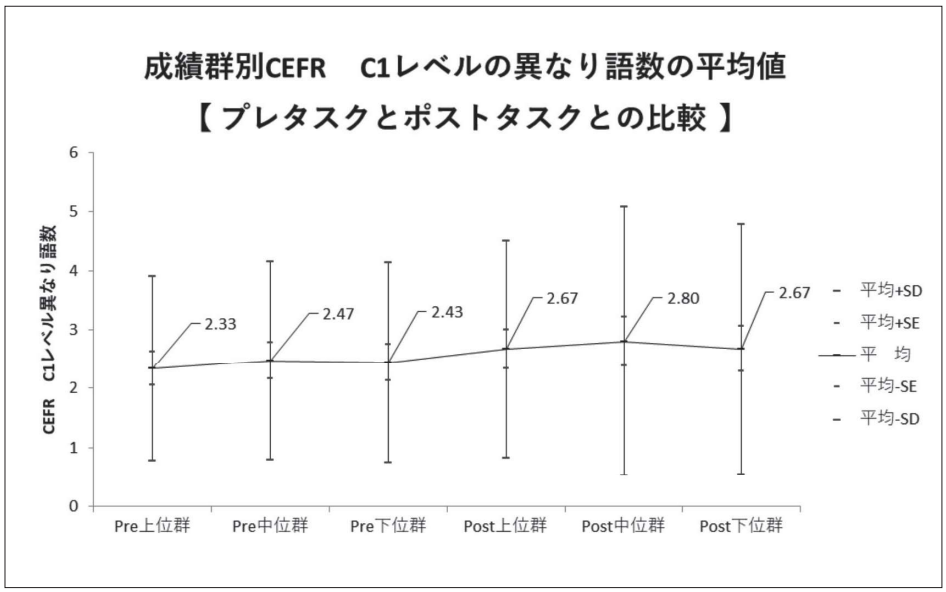
Note: *P<.05, **P<.01

3.7 成績群別CEFR C1レベル異なり語数(types)の平均値の差について

本節では，CEFR　C1レベルの異なり総語数の平均値の差について，調査を実施した(表25・図6)。分散分析の適用条件として，レーベン(Levene)等分散性検定の結果を確認したところ，有意確率が，p=.5121，つまり，5%水準で有意ではないので，「3群の分散は異なる」という帰無仮説が棄却された。つまり，「3群の分散は等分散である」ことになり，分散分析の適用が可能となった。そこで，一元配置の分散分析を行ったところ，3群の異なり語数の平均値の差は有意ではない(F(5,174))=.2654，n.s.)。その結果は表26である。

■表25: 成績群別CFFR C1レベルの異なり語数(平均値)

成績群	n	平均	標準偏差(SD)	平均-SD	平均+SD	標準誤差(SE)	平均-SE	平均+SE
Pre上位群	30	2.33	1.56	0.77	3.89	0.29	2.05	2.62
Pre中位群	30	2.47	1.68	0.79	4.14	0.31	2.16	2.77
Pre下位群	30	2.43	1.70	0.74	4.13	0.31	2.12	2.74
Post上位群	30	2.67	1.84	0.82	4.51	0.34	2.33	3.00
Post中位群	30	2.80	2.28	0.52	5.08	0.42	2.38	3.22
Post下位群	30	2.67	2.12	0.54	4.79	0.39	2.28	3.05



■図6：成績群別CEFR C1レベルの異なり語数の平均値比較

■表26：分散分析結果

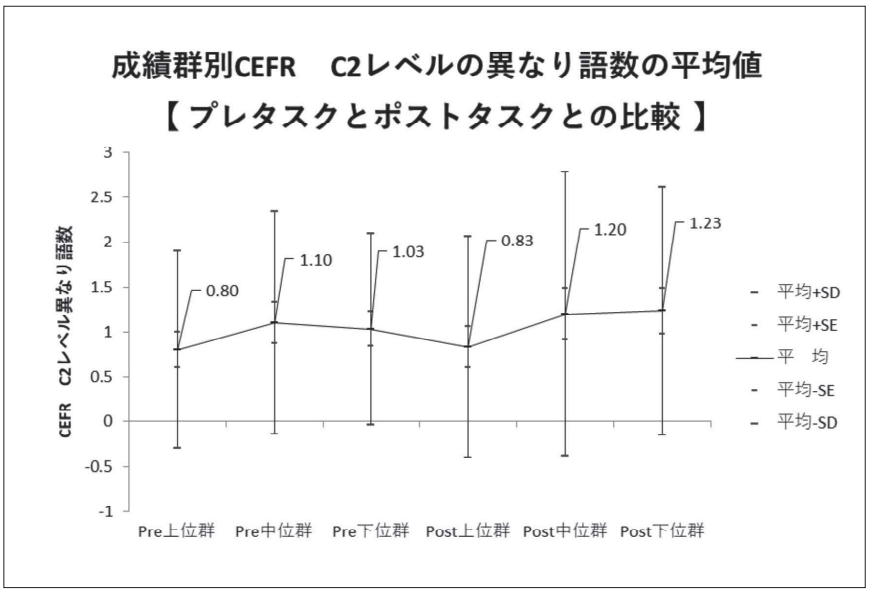
ソース	TypeⅢ平方和	自由度	平均平方	F値	P値
成績群	4.694	5	0.939	0.2654	0.9315
誤差	615.633	174	3.538		
全体	620.328	179			

3.8 成績群別CEFR C2レベル異なり語数(types)の平均値の差について

本節では，CEFR C2レベルの異なり総語数の平均値の差について，調査を実施した(表27・図7)。分散分析の適用条件として，レーベン(Levene)等分散性検定の結果を確認したところ，有意確率が，p=.4961，つまり，5%水準で有意ではないので，「3群の分散は異なる」という帰無仮説が棄却された。つまり，「3群の分散は等分散である」ことになり，分散分析の適用が可能となった。そこで，一元配置の分散分析を行ったところ，3群の異なり語数の平均値の差は有意ではなかった(F(5,174))=.6110，n.s.)。その結果は表28である。

■表27：成績群別CEFR C2レベル異なり語数(平均値)

成績群	n	平均	標準偏差(SD)	平均-SD	平均+SD	標準誤差(SE)	平均-SE	平均+SE
Pre上位群	30	0.80	1.10	-0.30	1.90	0.20	0.60	1.00
Pre中位群	30	1.10	1.24	-0.14	2.34	0.23	0.87	1.33
Pre下位群	30	1.03	1.07	-0.03	2.10	0.19	0.84	1.23
Post上位群	30	0.83	1.23	-0.40	2.07	0.23	0.61	1.06
Post中位群	30	1.20	1.58	-0.38	2.78	0.29	0.91	1.49
Post下位群	30	1.23	1.38	-0.15	2.62	0.25	0.98	1.49



■図7：成績群別CEFR C2レベルの異なり語数の平均値比較

■表28: 分散分析結果

ソース	TypeⅢ平方和	自由度	平均平方	F値	P値
成績群	5.000	5	1.000	0.6110	0.6916
誤差	284.800	174	1.637		
全体	289.800	179			

表29は，異なり語のCEFRレベル別語彙使用率を示した結果である。Unlistedの語彙数を除いたCEFR A1～C2までの6レベルの異なり語数の合計に占める各CEFRレベルの異なり語数の割合，すなわち異なり語数率を示している。例えば，CEFR　A1レベルに注目すると，ポストタスクにおいて成績群が上位になるほど，CEFR　A1レベルの語彙を使用しているのが分かる。

■表29: CEFRレベル別語彙使用率（異なり語数率:%）

			一人あたりの異なる総語数（Type）の平均語数：word					
			CEFR熟達度					
成績群	n	タスク	A1	A2	B1	B2	C1	C2
下位群	30	プレ	78.83	23.53	16.87	9.37	2.43	1.03
		ポスト	88.23	28.60	20.50	10.73	2.67	1.23
中位群	30	プレ	79.77	23.97	16.63	10.30	2.47	1.10
		ポスト	88.37	28.60	20.03	11.93	2.80	1.20
上位群	30	プレ	81.50	25.53	16.13	8.60	2.33	0.80
		ポスト	90.93	28.23	19.47	10.13	2.67	0.83

4 考察

前節で，分析結果を詳しく述べてきた。本節では，今回の実践研究の結果を踏まえて，研究課題を考察していく。プロダクト分析に関して，本実践研究で検討した仮説(H)とリサーチクエスチョン(RQ)は次のとおりであった。

H1　母語を使用した相互評価活動によるフィードバックを取り入れたピア・レスポンスを行うと，生徒は相手により理解されやすいエッセイを書こうとして，リライト活動でCEFRレベルA1・A2レベルの単語をより多く使うようになる。

RQ1　協働学習，特にピア・レスポンス後のリライト活動によって，生徒が使用する語彙に量的変化が見られるか。

RQ2　協働学習，特にピア・レスポンス後のリライト活動によって，生徒が使用する語彙に質的変化が見られるか。

以上のリサーチクエスチョンに対して答えるために，本実践研究でのプロダクト分析においては，まず，プレタスクとポストタスクで使用された異なり語数の総語数をt検定(対応あり)で分析した。これによって，プレタスクとポストタスクの一人あたりの異なり語数の差について調査した。プレタスクの異なり語数の総語数と，ポストタスクの異なり語数の総語数の差については，表8の結果より，有意な差が見られた。このことによって，参加者は，プレタスク時よりもポストタスク時の英文エッセイにおいて，より豊富な語彙を使用したことが推測される。しかし，この時点でのt検定(対応あり)による分析では，CEFRでの非リスト語(Unlisted)を含むため，プレタスクとポストタスクのCEFRレベル別の語彙使用数を比較するため，より詳細な分析を行う必要があった。そこで，CEFRのA1～C2の各レベルの異なり語数について，プレタスクとポストタスクを比較することになった。CEFR　A1・A2レベルでの異なり語数の平均値の差は，表9から有意であると解釈することができた。そのため，CEFR A1・A2レベルの語彙は，プレタスクよりもポストタスクにおいて，語彙数が増えていると解釈できるであろう。

CEFR　B1・B2の各レベルでの異なり語数の平均値の差については，表10から有意であると解釈することができた。CEFR　B1・B2レベルの語彙も，プレタスクよりもポストタスクにおいて，語彙数が増加しているが，この増加は有意であると考えられる。CEFR Cレベル(C1・C2)での異なり語数の平均値の差は，表11から有意でないと解釈することができ，CEFR Cレベル(C1・C2)の語彙は，プレタスク，ポストタスクにおいて，語彙数が増えてないと解釈することができる。

全体的な傾向を見ると，CEFR　AレベルとBレベルの語彙が増えている。各CEFRレベルの語彙の増加数で比較すると，A1[9.14語]，A2[4.80語]、B1[3.46語]，B2[1.51語]であった。特に，A1レベルの語彙の増加数はA2レベルの増加数の約2倍であった。このことから，全体的には，CEFR　A1レベルの語彙については，A2レベルの語彙よりも2倍程度多い9単語(異なり語)をリライト活動に用いていることが分かった。また，語彙に関連づけられたCEFRレベルがB1からC2へと上がるにつれて，その使用語彙数が少なくなっていることも特徴的であった。

延べ語数に関しても，CEFRの各レベル別に調査した。表12より，プレタスクとポストタスクにおいて，CEFR Aレベル(A1・A2)，Bレベル(B1・B2)の延べ語数の増加については有意な差が見られた。そのため，プレタスクからポストタスクで，一人あたりの延べ語数も増加していると見ることができる。また，CEFR　Cレベル(C1・C2)での延べ語数の平均値の差は，表12から有意でないと分かった。このため，CEFR　Cレベル(C1・C2)の語彙は，プレタスク，ポストタスクにおいて，語彙数が増えてないと解釈することができる。実際の増加数は，C1レベル[0.43語]、C2レベル[0.06語]であったことからも，増減に変化が見られなかったと言える。

次に，各CEFRレベルにおける一人あたり

の延べ語数の差に関して全体的な傾向を見た。A1[44.65語], A2[8.35], B1[5.33語], B2[2.33語]であった。A1レベルの語彙に関しては, A2レベルの語彙の約5倍多くリライト活動で使用されていた。

異なり語数において, CEFR Aレベル(A1とA2)とBレベル(B1とB2)の語彙の使用に関しては, プレタスクとポストタスクとで, 有意な差が見られたことが分かっている。そこで, どの成績群において有意な差があるのかについて詳しく調査するため, 英語学力によって参加者を, 各30名ずつの成績群(上位群・中位群・下位群)に分けて分析した。結果, どの成績群の参加者もCEFR A2レベルの異なり語において, プレタスクからポストタスクにかけて, A2レベルの語彙に有意な増加が見られた(プレ上位群とポスト上位群の差($p<.05$), プレ中位群とポスト中位群の差($p<.05$), プレ下位群とポスト下位群の差($p<.05$))。このことから, 成績群に関係なくすべての参加者が, 読み手により理解してもらえるように, CEFR A2レベルの語彙を使用していることが示唆された。他のCEFR BレベルとCレベルについては, A2のような有意な差が見られなかった。

5 結論

本実践研究では, 4技能のうちライティングに特化した取組を行った。そして, ライティングの力をどのように伸ばしていくのかという問いにも焦点を当て,「協働学習」を導入したライティングの授業デザインを考えた。今後の英語教育は「教師主体」の授業から,「学習者主体」の授業への質的転換が図られる。このため,「協働学習」の概念やアプローチが必要になるであろう。その意味で, ライティングと協働学習を組み合わせた実践研究は, 筆者にとっても, 学習者である生徒にとっても有意義なものであったと考える。

次期学習指導要領では, 育成したい資質・能力の観点から, 外国語教育(特に英語教育)においては, 外国語(特に英語)を用いて何ができるかという行動志向性をどのようなアプローチで高めるかが重要な課題であるとされている。そのため, 学習者の自律性や協働性を育むことをねらいとしたCEFRを授業活動の分析に取り入れたことは, 教師だけでなく, 学習者にとっても大きな意味があると考えている。

本実践研究において, ピア・レスポンスによる相互評価活動とその後のリライト活動が, エッセイにおける語彙使用に与える影響を調査するため, CEFRに関連づけられた語彙分析を行った。分析の結果, CEFR A2レベルにおいては, プレタスクとポストタスクで, すべての成績群において有意な増加が見られた。CEFR A2のレベルは,「基礎段階の言語使用者(Basic User)」(Council of Europe, 2001)であり, Bレベルの「自立した言語使用者(Independent User)」への一歩手前のレベルである。自分の力で書き換えができ, なおかつ, 自分が知っている語彙を多く使って書こうとしたときに, CEFR A2レベルの語彙を活用した生徒が多くいるということが示唆された。

次期学習指導要領では, 言語活動において,「支援」という言葉が使用されている。そして, この「支援」についても, 生徒が, より少ない支援で自律的に, 言語活動における課題を遂行する力が育成できるようにするために我々英語教師の適切な足場かけ(scaffolding)が求められている。生徒を「基礎段階の言語使用者」から「自立した言語使用者」として育成するために, 行動志向性の育成の観点から, 本実践研究で, 語彙レベル分析にCEFRの共通言語尺度を用いたのは, その試みの第一歩であった。今後も, 生徒の学習者自律性(learner autonomy)を高めるために, 英語教師として努力してきたい。

謝辞

この研究の機会をくださった公益財団法人日本英語検定協会, 選考委員の先生方, 特に池田央先生に心より感謝を申し上げます。また, この研究に参加してくださった愛媛大学附属高等学校の生徒のみなさまに深く御礼申し上げます。最後に, 遅々として進まなかった執筆を陰で支えてくれた家族にも感謝します。

参考文献（＊は引用文献）

＊ 阿部真・山西博之 (2013).「大学英語教育における協働的ライティング学習の可能性：グラウンデッド・セオリー・アプローチに基づいた分析の試み」『外国語教育メディア学会機関誌』第50号, pp. 93-117.

安藤昭一（編）(1991).『英語教育現代キーワード事典』大阪：増進堂.

＊ 馬場千秋 (2010).「第9章　ライティング指導でもとめられているもの」大学英語教育学会（監修）『英語教育学体系 第10巻　リーディングとライティングの理論と実践ー英語を主体的に「読む」・「書く」』(pp.117-134) 東京：大修館書店.

＊ Cho, M. & Reinders, R. (2013). The effects of aural input enhancement on L2 Acquisition. In J.M. Bergsleinthener, S., N. Frota, & J.K. Toshioka(Eds), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp.133-148). Honolulu: University of Hawaii, National Foreign Language Resource Center.

＊ Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

大学英語教育学会SLA研究会研究会（編）(2013).『第二言語習得と英語科教育法』東京：開拓社.

Dörnyei,Z. (2007). *Research Methods in Applied Linguistics,* Oxford: Oxford University Press.

＊ 江利川春雄 (2012).「第1章　協同学習の基本的な考え方」江利川春雄（編）『協同学習を取り入れた英語授業のすすめ』(pp.3-24) 東京：大修館書店.

＊ Ferris,D.(2003).Responding to Writing. In Kroll,B.(Ed.),*Exploring the Dynamics of Second Language Writing.* Cambridge: Cambridge University Press.

＊ Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34,* pp.906-911.

＊ 伏野久美子 (2010).「第10章　協働学習の理念と英語教育への応用」大学英語教育学会（監修）山岸信義・高橋貞雄・鈴木政浩（編）『英語教育学大系　第11巻　英語授業デザインー学習空間づくりの教授法と実践』(pp.94-102) 東京：大修館書店.

＊ Harrison, J. (2015). What is English Profile, in Harrison, J.& Barker (Eds.), *English Profile in Practice,* English Profile Studies volume5, Cambridge UCLES/Cambridge University Press.

＊ Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the Reference of the Common European Framework,* English Profile Studies volume1, Cambridge: UCLES/Cambridge University Press.

＊ 久山慎也 (2008).「高校生の自由英作文指導におけるピア・フィードバックの活用ープロセスの改善とライティング不安の軽減の視点から」『STEP BULLETIN』第20号, pp.166-178.

肥田野直・瀬谷正敏・大川信明（1961）.『心理教育　統計学』東京：培風館.

＊ 平林健治 (2016).「日本人上級英語学習者の自由英作文の語彙の豊かさに関する研究」『関東甲信越英語教育学会誌』30巻, pp.29-42.

＊ 広瀬和佳子 (2007).「教師フィーバックが日本語学習者の作文に与える影響ーコメントとカンファレンスの比較を中心に」『早稲田大学日本語教育研究センター紀要』第20号, pp.137-155.

＊ Hyland,K & Hyland,F.(Eds.)(2006).*Feedback in Second Language Writing: Contexts and Issues,* Cambridge: Cambridge University Press.

伊東治己（編）(2008).『アウトプット重視の英語授業』東京：教育出版.

＊ 岩中貴裕 (2014).「協同学習を取り入れた内容理解重視の授業ーそのリメディアル教育としての可能性ー」『四国英語教育学会紀要』第34号, pp.47-56.

＊ 鹿毛雅治 (2013).『学習意欲の理論ー動機づけの心理学』東京：金子書房.

＊ 神原克典 (2011).「ライティング活動におけるピア・レスポンスと教師フィードバックの効果」『STEP BULLETIN』第23号, pp.127-139.

小池生夫（編集主幹）(2003).『応用言語学事典』東京：研究社.

小池生夫（監修）SLA研究会（編）(1994).『第二言語習得研究に基づく最新の英語教育』東京：大修館書店.

＊ 小室敏明（編）(2001).『英語ライティング論ー書く能力と指導を科学する』東京：河源社.

＊ Matsuda,P.K. (2003). Second language writing in the twentieth century: A situated historical perspective. In Kroll,B.(Ed.),*Exploring the Dynamics of Second Language Writing.* Cambridge: Cambridge University Press.

三浦省五（監修）前田啓朗・山森光陽（編著）磯田貴道・廣森友人 (2004).『英語教師のための教育データ分析入門ー授業が変わるテスト・評価・研究ー』東京：大修館書店.

＊ 三宅なほみ・東京大学CoREF・河合塾（編）(2016)『協調学習とはー対話を通して理解を深めるアクティブラーニング型授業ー』京都：北大路書房.

＊ 三好徹明 (2016).「CEFRに準拠したリーディング補助教材開発のためのText inspectorによる英文テキスト難易度推定の試み」『四国英語教育学会紀要』第36号, pp.55-64.

＊ 文部科学省 (2009).『高等学校学習指導要領』Retrieved fromhttp://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2011/03/30/1304427_002.pdf　(2017年7月31日最終閲覧)

＊ 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書」文部科学省初等中等教育局.

＊ 文部科学省中央教育審議会 (2016).「幼稚園、小学校、中学校、高等学校及び特別支援学校の学習指導要領等の改善及び必要な方策等について（答申）」文部科学省初等中等教育局. Retrieved from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2017/01/10/1380902_0.pdf (2017年7月31日最終閲覧)

＊ 村上明(2013).「7章　CEFR基準特性と学習者英語　7･2【研究例･1】基準特性の抽出」投野由紀夫・金子朝子・杉浦正利・和泉絵美（編）『英語学習者コーパス活用ハンドブック』東京：大修館書店.

＊ 長崎睦子 (2017).「第2章　目標項目を目立たせようーインプット強化」鈴木渉（編）『実践例で学ぶ第二言語習得研究に基づく英語指導』(pp.13-26) 東京：大修館書店.

Nation,I.S.P.(2008).*Teaching Vocabulary: Strategies and Techniques,* Boston: Heinle Cengage Learning.

＊ 野上　泉 (2016).「高校生の英作文における「文章のまとまり」に焦点を当てたピア・フィードバック活動の効果」『STEP

参考文献（＊は引用文献）

BULLETIN』第28号, pp.85-104.

＊ Nunan,D. (Ed.)(1992). *Collaborative Language Learning and Teaching.* Cambridge: Cambridge University Press.

＊ 越智美江 (2015).「プロセス・ライティングのピア・フィードバック効果ーEFL学習者のピア・エバリュエーション分析ー」『四国英語教育学会紀要』第35号, pp.35-50.

＊ Philip,J., Adams,R.,&Iwashita,N.(2014). *Peer interaction and second language learning,* NY: Routledge.

Polio,C.and Friedman, D.A.(2017). *Understanding, Evaluating, and Conducting Second Language Writing Research,* NY: Routledge.

＊ Sato,M and Ballinger,S. (2016). Introduction: Understanding peer interaction. In Sato, M. and Ballinger, S. (Eds.) *Peer Interaction and Second Language Learning: Pedagogical potential and research agenda,* Amsterdam: John Benjamins.

＊ 佐藤礼子 (2012).「第11章　日本語教育における自己調整学習」自己調整学習研究会（編）『自己調整学習ー理論と実践の新たな展開へー』(pp.225-239) 京都：北大路書房.

佐藤雄大（2015）.『対話を用いた英語ライティング指導法』広島：渓水社.

＊ Storch, N.(2009). *The nature of pair interaction: Learner's interaction in an ESL class:its nature and impact on grammatical development.* Saarbrücken,Germany:VDM Verlag.

Storch, N.(2017). Sociocultural Theory in the L2 Classroom. In Loewen,S and Sato,M. (Eds.) *The Routledge Handbook of Instructed Second Language Instruction,* New York: Routledge.

＊ Schmidt,R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11(2),* pp.17-46.

竹内理・水本篤（編）(2014).『外国語教育ハンドブック【改訂版】研究手法のよりよい理解のために』東京：松柏社.

＊ 富永敦子 (2012).「文章表現授業における大学生のピア・レスポンス指向性の変化と要因の分析」『日本教育工学会論文誌』第36号, pp.301-311.

投野由紀夫（編）(2013).『CAN-DOリスト作成・活用　英語到達度指標CEFR-Jガイドブック（CD-ROM付』東京：大修館書店.

＊ 津田ひろみ (2013).『学習者の自律をめざす協働学習ー中学校英語授業における実践と分析ー』東京：ひつじ書房.

＊ Tsuchiya, M. (2015). Collaborative Learning for Less Proficient Learners in University English Education: Combatting Demotivating Factors. *ARELE 26,* pp.381-395.

＊ 山本恭子 (2010).「第2言語ライティング学習時に協働作業によるピアレスポンスが生む創造性」『STEP BULLETIN』第22号, pp.128-144.

＊ 吉田達弘 (2001).「社会文化的アプローチによる英語教育研究の再検討：「獲得」から「アプロプリエーション」へ」『兵庫教育大学言語表現学会言語表現研究』、第17号, pp.41-51.

吉島茂・大橋理枝（編）(2004).『外国語教育Ⅱ＜追補版＞ー外国語の学習、教授、評価のためのヨーロッパ共通参照枠ー』東京：朝日出版社.

＊ Zimmerman,B.J., Bonner,S.,& Kovach, R. (1996). Goal 1, Understanding the principles of self-regulated learning. In B.J. Zimmerman,S. Bonner & R. Kovach. (Eds.), *Developing self-regulated learners: Beyond achievement to self-efficacy.* American Psychological Association.

資料1：ピア・フィードバックで使用したワークシート

ピア・フィードバックをおこないましょう。（英語で書ける人は、英語で書いてください。）

1　友達のエッセイのどんな点が一番好きかを書きましょう。（内容・表現方法など）

2　自分自身の言葉で、このエッセイは何について書いてあるのかを述べて。簡単に記録してください。

3　読み手（あなた）が理解しやすいように、このエッセイのどこをさらに詳しく言えばよいかと述べましょう。

4　一番、同意できる意見を選んで書きましょう。

5　パートナーへの短いコメントを書きましょう。コメントには、どうすればエッセイの質が向上するかを説明してください。また、その理由も書きましょう。

Dear（　　）　Title（　　）

Name	Messages to the writer

第30回 研究助成

C 調査部門 ● 報告Ⅰ ● 英語教育関連の調査・アンケートの実施と分析

LFCにおけるcore項目の学習：日本人英語学習者に関する音響分析

研究者：神奈川県／慶應義塾高等学校　教諭 北川 彩

《研究助言者：吉田 研作》

概要 本調査の目的は，日本人英語学習者が共通語としての英語（English as a Lingua Franca: ELF）使用者として必要な能力を身につけるにあたり，学習や指導が必要とされる発音項目における学習の困難度について，音響分析の手法を用いて調査することである。清水（2011）に基づいて学習や指導が必要と思われる項目をcoreとし，当該項目に関して3人の英語母語話者と24人の日本人英語学習者のデータを分析した結果，/t/の気音，/v l ɪ/の調音，狭いフォーカスと文構造などに依存して決定される音調核の位置の学習の困難度が高いことが分かった。また，語のアクセントについても，全体的には困難度が高いわけではないが，弱音節の短さやあいまいな母音の質を生成することの困難度は高かった。一方で，/w/の調音，上昇調を核音調として使うこと，広いフォーカスの場合に音調核を正しく置くことは困難度が低い項目であることが示された。こうした結果は，ELF時代における発音学習や指導の方向性について新たな示唆を与えるものと期待される。

1 はじめに

共通語としての英語（English as a Lingua Franca：ELF）が広まった現代において，英語母語話者の定義は多様化し，母語話者を介さない場面における英語の使用も一般化した。それに伴い，学習者が英語を学ぶ目的も多岐に渡るようになった。英語学習者の発音モデルとして英語母語話者の発音を唯一のものとするという考えに対して議論が起こり始めたのはそうした流れの中においてであり，例えばSeidlhofer（2011）は英語学習の目標は学習者自身の目標によって定められるべきものと述べている。

清水（2011）が示しているように，音声学の教科書として世界的に広く使用されてきた*Gimson's Pronunciation of English*の最新版（Cruttenden, 2014）においても学習を優先すべき発音項目をまとめたモデルとしてInternational Englishが提案されている。またJenkins（2000）も，コミュニケーションの阻害になるかどうかという視点から発音項目をcoreとnon-coreに分類し，Lingua Franca Core（LFC）と呼ぶ発音の指針を提示した。こうした発音が学習者の最終的な目標となりうるかどうかについての議論はあるものの（Alptekin, 2007; Cogo, 2012; Dewey, 2013; Kuo, 2006, 2007; Sewell, 2013; Sowden, 2012; Sung 2013），学習過程において目指すべき発音学習と指導の方向性を新たな視点から示すという試みは意義深い。

第一言語（L1）の音声・音韻上の特徴が目標言語の学習に影響を及ぼすという前提に立ち，Jenkins（2000）のLFCに準拠しながら，L1ごとに学習すべき発音のガイドラインを具体的に提案しているのがWalker（2010）である。Walkerは，ELF使用者として最低限の通じやすさを保つ上で必要な能力を身につけるためにL1をうまく活用することを提案し，そこに記された10の言語において

も，L1を出発点としてL1の特徴を活用できるものはnon-coreとしながら，指導上coreとなる項目の提示を行っている。そこで記されている日本人英語学習者のためのガイドラインについて，さらに詳述されたものが清水(2011)である。

2 本調査の背景

2.1 清水(2011)によるガイドライン試案

清水(2011)はガイドラインの中で，coreとnon-coreという表現を直接使用しているわけではない。そのため，北川(2017)ではLFCとの関連を考えながら，日本語の特徴を代用しても通じやすさに問題ないとしている項目をnon-core，それ以外をcoreと分類し，coreとされた項目を中心に表1のようにまとめた。日本人英語学習者に向けた清水(2011)には，LFCの分類と異なる点もいくつか見られる。北川(2017)における解釈に沿って簡単にまとめると，異なる分類を示している項目は，(1)子音連結，(2)母音の質，(3)語強勢の3項目である。(1)清水は，日本人英語学習者においては母音の添加が問題になることを指摘し，特に母音の添加は避ける指導が必要としている。(2)清水は，/ɪ/の質の学習を勧めており，また/æ/と/u:/の質についても調音の仕方を説明しているので，指導が必要なcoreな項目と考えていると見なした。(3)清水は，語強勢の有無と母音の質には表裏一体の関係がある，語強勢の誤りが通じやすさに影響を及ぼさないと考えることには疑問が残る，coreとする音調核の位置は語強勢に基づくということを指摘し，母音を強く長く発音し語強勢を置くことを必要としている。いずれもLFCではnon-coreとされているが，清水ではcoreとされていると思われる項目である。判断における詳細な記述は北川を参照されたい。

2.2 調査課題

本調査の目的は，清水(2011)のガイドライン試案において指導が必要とされている項目について，音響分析の手法を用いて日本人英語学習者にとっての学習の困難度を調査することである。学習の難しさを調査することにより，ELF使用者を目指す日本人英語学習者にはどのような発音学習と指導が必要なのか，方向性を検討することが可能になる。本調査はその試みとして，清水によって音環境に関係なくcoreとされている分節音の調音に関する項目と，プロソディに関するcore項目の学習困難度を示すことを課題とした。

3 調査方法

3.1 被験者

本調査の参加者は，日本人英語学習者に対して日常的に英語を教えている英語母語話者3人(アメリカ英語話者2人とオーストラリア英語話者1人)と，全体の発音レベルによって8人ずつの上位・中位・下位に分けられた24人の日本人英語学習者である。24人は，できるだけ様々な発音レベルを持つ日本人英語学習者から調査課題とした学習困難度を捉えるために，データを収集した53人の中から英語母語話者1人による聴覚実験の結果に基づき選ばれた被験者である(注1：219頁)。被験者は全て男性である。

3.2 録音とデータの収集

デジタル録音機R-09とコンデンサーマイクECM-MS957を用い，録音室にて1人ずつ録音を行った。日本人英語学習者に関しては，言語材料を事前に配布し，内容に関して不明な点がないか確認した上で録音を行った。その際，録音機は44.1 kHz, 16-bitでデジタル化するように設定し，被験者ごとに音量調節をしてから録音を開始した。

3.3 言語材料

調査対象としたのは，清水(2011)の中で/u:/を除く，音環境に依存しない分節音に関する11項目(/p t k/の気音，/f v r l w ɪ æ ɜ:/の調音)と，プロソディに関する5項目[アクセント，イントネーション句(Intonation phrases: IPs)の区切り，

音調核の位置，上昇の核音調，下降の核音調］である。分節音に関する項目は単語を1回ずつ読み上げたもの，プロソディに関する項目は，Wells（2006）に掲載された *Towels, Getting Breakfast, Books, Cornwall* の4つの分析用文章を1回ずつ読み上げたものである。誤った場合は，何度読み上げ直してもよいとした。詳細な言語材料は表2に示す。

■表1：清水（2011）による日本人英語学習者のためのガイドライン試案（北川，2017〔一部修正〕）

項目			Core	Non-core
子音	破裂音	/p/	気音	
		/t/	気音 [tʃ]を代用しない</i: ɪ u: ʊ/の前>	弾き音化
		/k/	気音	
		/b/	摩擦音化を避ける<語中>	
		/d/		弾き音化
		/g/	摩擦音化を避ける<語中> [ŋ]を代用しない<語中>	
	摩擦音	/f/	調音位置・様式	
		/θ/	[ʃ]を代用しない</i: ɪ/の前>	調音位置
		/s/	[ʃ]を代用しない</i: ɪ/の前>	
		/h/	[ç ɸ]を代用しない</i: ɪ u: ʊ/の前>	
		/v/	調音位置・様式 [b]を代用しない	
		/ð/	[dʒ ʒ]を代用しない</i: ɪ/の前>	調音位置
		/z/	[dʒ ʒ]を代用しない</i: ɪ/の前>	
	接近音	/l/	[l]の調音様式・調音位置	
		/r/	調音様式・調音位置	
		/w/	調音様式・調音位置	
	鼻音	/n/	[N]を代用しない<語末>	
	子音連結		母音の添加を避ける<語中・語間の 3子音連続における真ん中の/t d/も含めて>	
母音			/ɪ æ ɜ: u:/の質[1]	
アクセント			語強勢	
リズム				弱形 強勢拍リズム
イントネーション			区切り 音調核の位置 下降と上昇の核音調パターン	下降上昇調・平板調・上がり続ける上昇調の核音調
連続発音における音変化				脱落や同化

Note: [1]．清水（2011）では，/ɜ:/との関係から/u:/の調音もcoreと見なしていると思われるが，本調査ではL1をできる限り活かすというLFCの立場をとり，/ɜ:/をcoreとし，/u:/は日本語の「ウー」の代用で構わないと考えた。そのため，/u:/の調音は本調査にも含めなかった。北川（2017）参照。

■表2: 言語材料

分析項目	言語材料	
/p/気音	pen / puff / pat / paid / park / pub	
/t/気音	take / tell / touch / tap / ten	
/k/気音	cart / cut / cake / cap / cat / card	
/f/調音	fit / full / feet / foot / fork / fool / feed / food	
/v/調音	vest / voice / visit / very / vast	
/r/調音	rid / road / read / rude / rice / red / rat / ring	
/l/調音	loan / lake / look / leg / lack / lip / lead	
/w/調音	wet / week / win	
/ɪ/調音	sit / dip / fit / ship / this / kick / tip / pick / zip / chip / bit / thick / kid	
/æ/調音	cap / dash / that / gap / pat / gas / tap / Jack/ cat / chat / bad	
/ɜ:/調音	shirt / church / birth / purse / third / dirt	
アクセント[1]	happy / again / replace / excuse / breakfast / seventh1 / seventh2 / twenty1 / number / return /essay / suppose / weather	
区切り[2]	入れる	If you're staying more than one night | and are happy to use your towels again … Use the lift | over there. Well the best thing we did | was to go to the Eden Project. It's a kind of museum | of ecology.
	入れない	In the Panorama restaurant, (x) sir. Twenty-seventh floor, (x) sir. Use lift number five, (x) sir. Are you going to return those books of mine (x) you borrowed? The ones about biology (x) and language.
音調核の位置[3]	広い	bath / help / excuse / breakfast / that / there / see / language / days / essay / back / Cornwall / weather / that / ecology / fascinating
	狭い	replace / fourth / seventh2 / which / I or remember / those / been / marvelous / do / children
	その他[4]	towel / restaurant / floor / five / books1 / suppose / away1 / was / rained / Eden / theme
上昇調の核音調	books / days / OK / away1	
下昇調の核音調	towel / bath / help / breakfast / restaurant / that / floor / there / fourth / there / fourth / ah / five / seventh2 / thanks / hey / which / language / Oh / those / why / essay / back / been / Cornwall / was / marvelous / weather / rained / do / Eden / that / ecology / fascinating / theme / children	

Note: 1.付与された数字は，文章中で当該の語が登場する順番を表す。例えば，seventh1とは最初に登場するseventhということを示す。2.区切るべきとされている箇所は｜の記号で示し，区切らないとされている箇所は(x)で示した。3.付与されている数字はアクセントの項目と同様に，当該の語が文章中に登場する順番を意味にする。1つのIPの中で音調核の位置に2つ候補があるものは，orで表した。4.その他とは文構造などに関するもので，複合名詞(towel, Eden, theme)，呼格(restaurant, floor, five)，関係詞節(books1)，代用形(suppose)，時間表現の副詞(away1, rained)，why + to be(was)が含まれる。

3.4 音響分析

データの音響分析では，音響分析ソフトウェアのPraat（Boersma & Weenink, 2011, 2015）を使用した。スペクトログラム，波形，インテンシティの動きを観察しながらセグメンテーションを行って分析対象とした箇所を特定した後，表3に示す音響分析を行った。

■表3: 音響分析

調査項目	音響分析の項目	変数の単位
/p t k/の気音	Voice Onset Time（VOT）の長さ	ms
/f v/の調音	スペクトルピーク（Jongman, Wayland, & Wong, 2000）[1]	Hz
	2次のモーメント（M2）（Jongman et al., 2000）[1]	Hz
	3次のモーメント（M3）（Jongman et al., 2000）[1]	
	標準化された振幅（Jongman et al., 2000）[2]	
/l r/の調音/l r/	第三フォルマント（F3）（Ladefoged, 2003; Saito & Lyster, 2011）	mel[9]
/l/	調音区間の長さ	ms
/w/の調音	前舌短母音に対するフォルマント遷移時間の比[3]	
/ɪ æ ɜ:/の調音	第一フォルマント（F1）	mel[9]
	第二フォルマント（F2）	mel[9]
アクセント	弱音節と強音節の母音の調音区間の長さ[4]	ms
	弱音節と先行・後続する強音節のインテンシティの差[5]	dB
	弱音節と先行・後続する強音節の最大ピッチの差[6]	semitone[9]
	弱音節の母音のF1とF2[7]	mel[9]
区切り[8]	入れるべき箇所に入っているか	
	入れない箇所に入っていないか	
音調核の位置[8]	広いフォーカスにおいて置かれるべき箇所に置かれているか	
	狭いフォーカスにおいて置かれる箇所に置かれているか	
	文構造などによって決まる場合に置かる箇所に置かれているか	
下降調[8]	陳述文で下降しているか	
	wh疑問文で下降しているか	
上昇調[8]	上昇しているか	

Note: 1.22.05 kHzへの再標本化と高域強調を行い，スペクトルピークと2次のモーメントと3次のモーメントは，40 ms区間でハミングウィンドウの窓関数を掛けて行ったFFTとLPCから測定を行った。スペクトルピークは破裂音中央の40 ms，モーメントについては，破裂音の開始から40 msと破裂音と母音にまたがる40 msで計測した。2.後続する母音において最もインテンシティの強い箇所を中心にした3周期間の平均値と，摩擦音調音区間におけるインテンシティの平均値の差を計算した。3./w/と/u/の音響上の区別のひとつは，前者の方がF2におけるフォルマントの遷移にかかる時間が短いことであるため（Kent & Read, 2002），高いF2を持つ/i: ɪ e/が後続する場合を分析対象とし，F2の遷移がなだらかになる直前までの長さを計測した。ただし，Kent and Read（2002）は遷移時間には発話速度の影響があることも指摘しているため，後続する短母音の長さをwetとwinにおいて記録し，それらに対する相対的な時間長を統計分析に用いた。4.happyとseventh2の強音節，again, excuse, supposeの弱音節を分析対象とした。5.happy, again, replace, excuse, breakfast, seventh2, essay, suppose, weatherの弱音節とその前後いずれかの強音節を計測した。6.核音調が置かれないことを考慮して，seventh1，twenty2, number, seventh2の弱音節と前後いずれかの強音節を測定した。7.again, breakfast, suppose, weatherの弱音節において測った。8.Wells（2006）の示すモデルと比較し，一致の場合は1点，不一致の場合は0点と得点を与え，統計分析のための変数とした。その際，区切りの項目の呼格に関する3調査文（In the Panorama restaurant, sir. // Twenty-seventh floor, sir. // Use lift number five, sir.），音調核の項目の複合名詞に関する3調査語（towel rail, Eden project, theme park），呼格に関する3調査語（restaurant, floor, five），時間表現の副詞に関する2調査語（away1, rained）は各合計1点となるよう計算した。変数の最大値は，区切り入れるが4，区切り入れないが3，広いフォーカスにおける音調核の位置が16，狭いフォーカスにおける音調核の位置が10，その他の音調核の位置が6，陳述文の下降調が27，wh疑問文の下降調が7，上昇調が4である。下降調については，当該の核音調が使われる主な文のタイプに陳述文とwh疑問文があるため，2つに分けて分析した。上昇調の調査文はyes-no疑問文3つと確認1つである。9.クラスター分析においては，知覚特性を表す尺度として，母音と/r l/のHzはmelへ，ピッチのHzはsemitoneへ変換したものを用いた。

3.5 統計分析

英語母語話者の発音特性に近い日本人英語学習者の割合によって学習困難度を定義し，分析を行うこととした。そのため，音響分析にて得られた数値を標準得点に変換してクラスター分析を行い，被験者を学習の程度によって分類した。分類方法には解釈しやすいWard法を用い（足立，2006），距離の測定にはユーグリット平方距離の測定方法を用いた階層的クラスター分析を実行し，樹形図の結果から最大3段階に分類できるようクラスター分けをした。その際，英語母語話者3人が1つのクラスターにまとまることをクラスター分けの第一の基準とし，英語母語話者と同じクラスターに分類されている被験者の学習段階を1，それ以外を2と3に分け，3を学習が最も初期に近い段階，2は1と3の中間に相当する学習段階と位置づけた。ただし，クラスタリングにおいて英語母語話者3人が2人と1人に枝分かれした場合は，多数派の方に近づくことを本調査における学習と定義し，前者を学習段階1，後者を学習段階2とした。また，最大3段階に分けることとしたため，英語母語話者のまとまり方によっては，3段階ではなく2段階に定めることも許容している。

下記に示す2つの場合は，クラスター分析に含めなかった。ひとつは，音響分析上の問題に関するものである。/f v/のスペクトルの分析において摩擦音の区間が40 ms以下の場合は音響分析を行わず，アクセントの分析において音が脱落している場合は，長さは0 msとしたがその他の分析は行わなかった。従って，これらに該当するデータはクラスター分析にも含まれていない。もうひとつは，代用に関するものである。/v r l/の音響分析では，まずスペクトラム，波形を観察し，聴覚印象にも基づきながら/v/から[b]への代用，/r l/の[ɾ]への代用が行われていないかを確認し，代用が行われていると判断した語が調査語の中で半分以上に相当する場合は残りのデータもクラスター分析には含めなかった。そしてそのような場合は，学習がまだ初期段階に近いと見なして学習段階3とした。

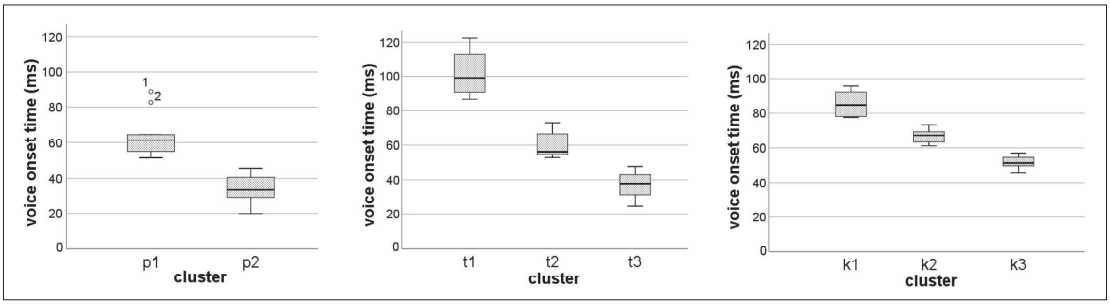
4 結果と考察

クラスター分析の結果に基づいて1から3，あるいは1か2のそれぞれの学習段階に分類された人数分布，形成されたクラスターごとの記述統計，その特徴を表す図表を示す。箱ひげ図には外れ値となった被験者が示されているが，1，2，3は英語母語話者である。関連している項目はひとつにまとめ，以下には，/p t k/の気音，/f v/の調音，/r l/の調音，/w/の調音，/ɪ æ ɜ:/の調音，アクセント，イントネーションの順に記す。

4.1 /p t k/気音

/p/では2つのクラスターへ，/t k/では3つのクラスターへと分類した。どの破裂音においても学習段階1は3人の英語母語話者を含み，最もVOTが長いことがその特徴であるため，長さが長くなるほど学習が進むと定義して学習段階を定めた。図1が示す箱ひげ図はこの傾向を視覚的に表している。

表4は，学習段階別の人数と，VOTの平均値，SDをまとめたものである。学習段階1に分類された人数によって3つの破裂音を比較すると，学習の困難度に差異があることが分かる。学習段



■図1：/p t k/のVOTの長さ

階1の日本人英語学習者が最も少なかったのは/t/の1人で，/p/と/k/では7人が学習段階1とされた。また，3段階に分かれた/t/と/k/を比べると，学習段階3に分類されたのは/t/では14人，/k/では7人であった。こうした人数の結果から考えると，3つの破裂音の中では/t/の困難度が最も高いと言える。/t/の難しさは清水(2008)の研究においても示唆されるところである。清水の測定では日本語の/t/のVOTは30 msと3つの中では最も短い。一方で，英語の/t/のVOTはLisker and Abramson(1964)や清水(2008)の研究では/k/に次いで長く，本調査では，図1の箱ひげ図を見ても分かるように最も長かった。このことが，短い/t/の気音を持つ日本人英語学習者にとっての英語の/t/の気音の難しさに起因していると考えられる。

■表4: /p t k/の学習段階別人数分布と記述統計

	学習段階	*n*	全体発音レベルごとの*n*				平均	SD
			母語	上位	中位	下位	VOT	VOT
/p/	1	10	3	4	2	1	63.86	12.43
	2	17	0	4	6	7	34.44	8.13
/t/	1	4	3	1	0	0	101.80	15.35
	2	9	0	5	3	1	60.42	7.56
	3	14	0	2	5	7	37.53	7.29
/k/	1	10	3	4	3	0	85.37	6.87
	2	10	0	4	2	4	66.92	4.14
	3	7	0	0	3	4	51.71	3.79

Note: VOTの値はmsが単位である。

4.2 /f v/調音

/f/は2つの学習段階，/v/は3つの学習段階に分類された。ただし，/v/の場合，音響分析にて[b]の代用があると判断された日本人英語学習者が14人，摩擦音の区間が短く分析できなかった日本人英語学習者が1人おり，クラスター分析は12人で行った。そのため，/v/において学習段階3とされた被験者は全て[b]への代用があった場合であり，数値が得られたのは/f/も/v/も学習段階2までである。図2の箱ひげ図は，スペクトルの形状を表す変数として扱ったスペクトルピーク，2次のモーメント(M2)，3次のモーメント(M3)と，音の強さに関する変数である標準化された振幅の結果を表したものである。

表5は学習段階別の人数を表し，表6は学習段階

■表5: /f v/の学習段階別人数分布

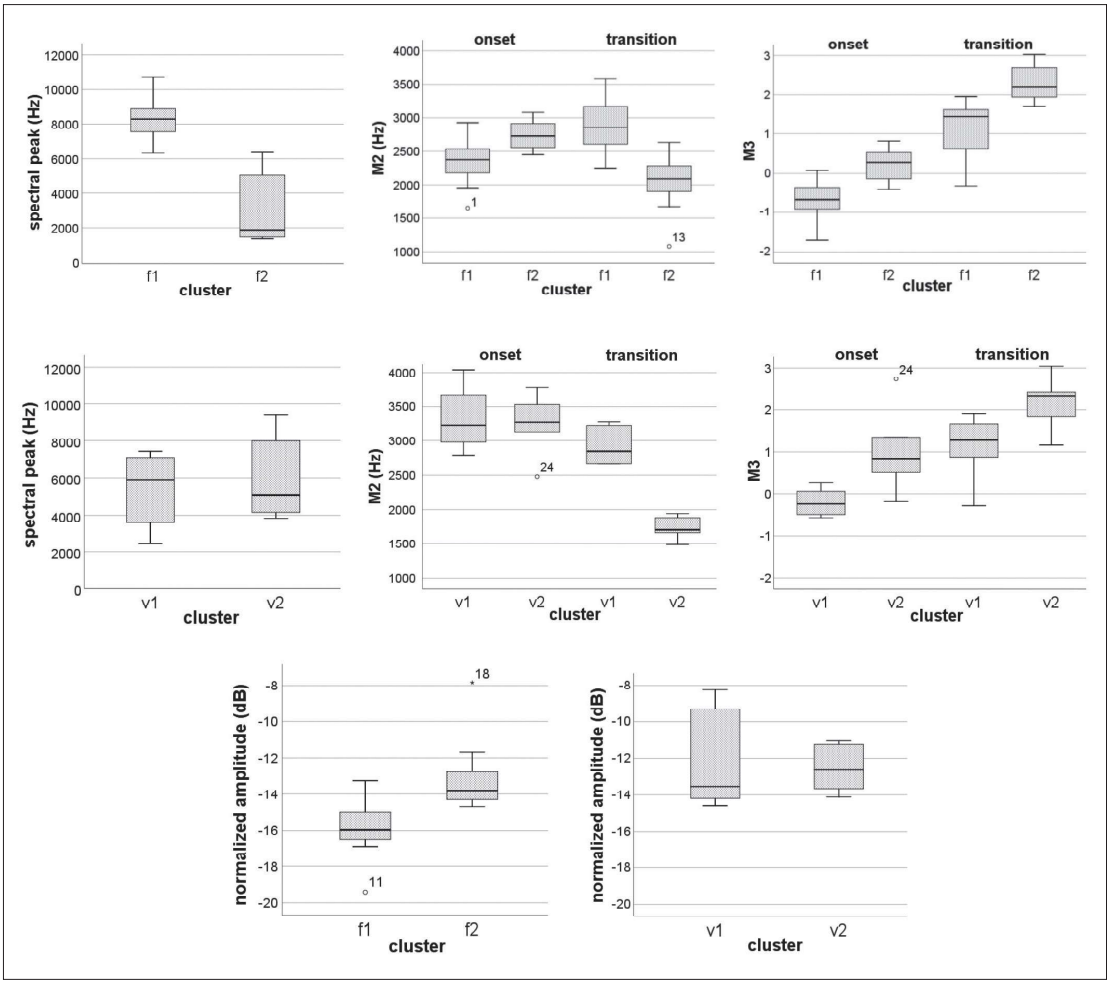
	学習段階	*n*	全体発音レベルごとの*n*			
			母語	上位	中位	下位
/f/	1	15	3	8	3	1
	2	12	0	0	5	7
/v/[1]	1	6	3	3	0	0
	2	6	0	3	1	2
	3	14	0	1(1)	7(7)	6(6)

Note: カッコ内の人数は，調査語の半分以上に関して[b]への代用があり，学習段階3とした被験者数を表す。[1]．摩擦音の区間が短く分析できなかった日本人学習者が1人いたため、分析対象となった被験者数は26名である。

■表6: /f v/の記述統計

	学習段階	平均						SD					
		ピーク	振幅	M2:1	M2:2	M3:1	M3:2	ピーク	振幅	M2:1	M2:2	M3:1	M3:2
/f/	1	8287.97	−15.82	2352.93	2888.75	−0.66	1.13	1308.10	1.47	308.35	382.45	0.45	0.66
	2	3131.58	−13.15	2734.39	2037.74	0.22	2.28	2020.19	1.88	215.22	400.86	0.40	0.43
/v/	1	5399.88	−12.24	3322.66	2919.98	−0.21	1.12	2025.75	2.74	458.86	282.90	0.32	0.78
	2	5930.27	−12.53	3243.58	1728.96	1.02	2.19	2306.90	1.36	445.11	164.12	0.98	0.63
	3	–	–	–	–	–	–	–	–	–	–	–	–

Note: M2:1 = 摩擦音始まり40 ms のM2, M2:2 = 摩擦音と母音にまたがる40 ms のM2, M3:1 = 摩擦音始まり40 msのM3, M3:2 = 摩擦音と母音にまたがる40 ms のM3。

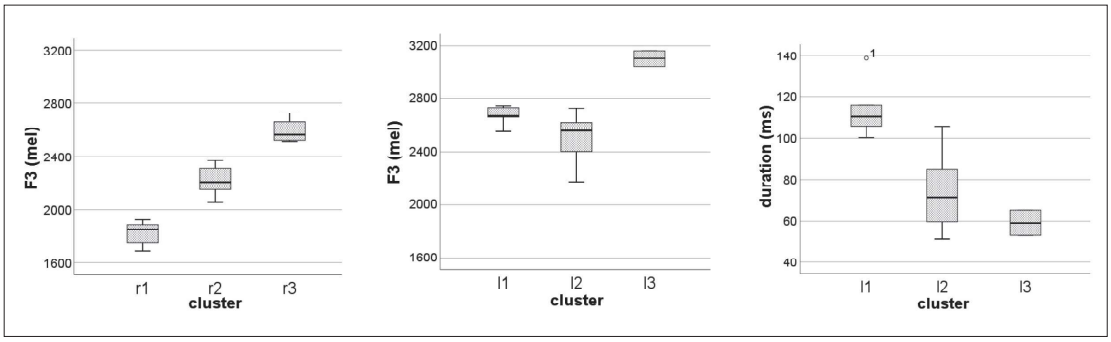


■図2:/f v/のスペクトルピーク(1段目左と2段目左);/f v/の標準偏差(M2)(1段目中央と2段目中央); /f v/の歪度(M3)(1段目右と2段目右);/f v/の標準化された振幅(3段目)

ごとの平均値とSDをまとめたものである。2つの摩擦音を比較すると，学習段階1とされた日本人英語学習者は/f/では12人，/v/では3人であり，/v/については半数以上の日本人英語学習者において[b]への代用が観察された。代用については東後・御園(2009)も指摘するところである。また，全体的な発音レベルが上位の被験者は全て/f/の学習段階1とされている一方で，当該被験者のうち/v/において学習段階1とされたのは3人のみであったことも/v/の学習の困難さを示している。ただし，図2に示した箱ひげ図からも視覚的に分かるように，学習段階1と学習段階2の違いを比較した場合，スペクトルの標準偏差と歪度を示す2次のモーメントと3次のモーメントは/f/と/v/どちらにおいても同様の傾向を表しているものの，スペクトルピークと，母音に対する強さの差を表した標準化された振幅については/v/の方が学習段階間の差が小さい。こうした結果からは/v/の困難度の方が低いのかもしれないという解釈も生ずるが，/v/の方がクラスター分析の人数が12人と少なく，それが/f/よりも細かいクラスター分けの結果につながった可能性があるということは考慮しなければならない。従って，上記に述べた全体の人数分布の傾向を基に，本調査では/v/の困難度の方が高いのではないかと考えた。

4.3 /r l/調音

音響分析の段階で，/r/では3人の日本人英語学習者について，/l/では6人の日本人英語学習者について日本語の[ɾ]への代用が半分以上の調査語にあると判断され，クラスター分析はそれぞれ24人，21人の被験者で行った。分析の結果，英語母語話者3人を含む学習段階1と，その近さから定義された学習段階2と3のクラスターが得られた。図3は，/r/と/l/のF3の値と，/l/の調音区間の長さを表した箱ひげ図である。



■図3：/r l/のF3(左と中央)；/l/の長さ(右)

■表7：/r l/の学習段階別人数分布と記述統計

	学習段階	*n*	全体発音レベルごとの*n*				平均		SD	
			母語	上位	中位	下位	F3	長さ	F3	長さ
/r/	1	10	3	5	2	0	1826.21	–	83.62	–
	2	4	0	0	2	2	2214.92	–	101.21	–
	3	13	0	3	4(1)	6(2)	2590.81	–	95.92	–
/l/	1	5	3	0	1	1	2670.66	114.29	76.64	14.80
	2	14	0	4	7	3	2504.57	74.07	181.09	16.44
	3	8	0	4(2)	0	4(4)	3101.56	59.14	85.63	8.69

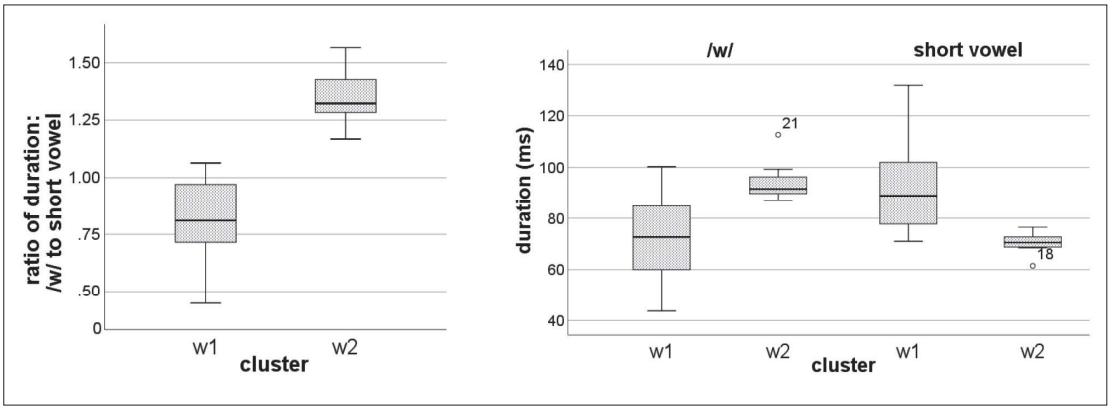
Note: カッコ内の数値は，調査語の半分以上に関して日本語の[ɾ]への代用があるとして学習段階3と分類した被験者数を表す。F3の値はmelである。

表7は学習段階の結果と，学習段階ごとの/r/と/l/のF3，/l/の長さの平均値とSDを表したものである。2つの音を比較すると，学習段階1に分類された日本人英語学習者が/r/では7人，/l/では2人と前者の方が多く，また日本語[ɾ]への代用があると判断された被験者数も/l/の方が多かった。その結果を考慮すると，/r/の困難度の方が低いと考えられる。/r/の方が学習しやすいということはAoyama, Flege, Guion, Akahane-Yamada, and Yamada (2004) も指摘している。ただし，Aoyama et al.の実験では，時間を空けて行った実験のうち1回目の実験では/r/よりも/l/のほうでより良い生成が報告されているため，学習段階の始まりについての検証はもう少し必要である。本調査においては，日本語の[ɾ]への代用があると判断された被験者を主とする/l/の学習段階3を除いて，/l/の学習段階2と/r/の学習段階3のデータを観察すると，両者のF3の値が似通っており（図3），このことは学習があまり進んでいない段階では/r/と/l/の区別がなされていない可能性を示唆する。

4.4 /w/調音

3人の英語母語話者を学習段階1として，2つの学習段階を定めた。変数である/w/のフォルマント遷移にかかった時間の後続する短母音の調音時間に対する比率とともに，/w/と短母音の長さを各学習段階で箱ひげ図によって表したものが図4である。

表8は2つの学習段階の人数分布と，/w/と短母音の長さ比，/w/の長さ，短母音の長さの平均値とSDを表したものである。学習段階1とされたクラスターと学習段階2とされたクラスターについて数値を比較すると，前者の方が/w/の長さが短く，比率も小さい。清水（2011）が「日本語の「ウ」を使わずに（p.58）」と指摘しているように，日本人英語学習者にとっての/w/の学習が/u/との比較において始まるとすれば，学習段階1が示すこの特徴は/w/の学習を反映したものであると捉えることができる。そして24人中17人の日本人英語学習者が学習段階1とされた結果を見れば，困難度は低い傾向にあると言えるだろう。



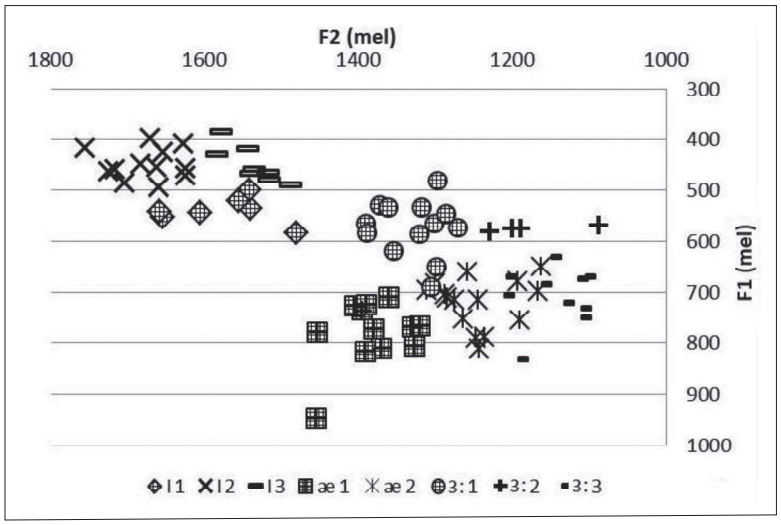
■図4：/w/のフォルマント遷移にかかる時間比（左）；/w/と後続短母音の時間長（右）

■表8：/w/の学習段階別人数分布と記述統計

学習段階	n	全体発音レベルごとのn				平均			SD		
		母語	上位	中位	下位	長さ比	/w/長	短母音長	長さ比	/w/長	短母音長
1	20	3	5	5	7	0.81	73.35	92.90	0.19	15.45	24.48
2	7	0	3	3	1	1.35	94.81	70.21	0.13	8.78	4.72

4.5 母音 /ɪ æ ɜː/ 調音

クラスター分析の結果，/æ/ と /ɜː/ では英語母語話者3人を学習段階1とした2段階と3段階にそれぞれ分類したが，/ɪ/ では英語母語話者が2人と1人のクラスターに分かれたため，前者を学習段階1，後者を学習段階2とし，残りの1つクラスターを学習段階3とした。図5は学習段階別に3つの母音のF1とF2を散布図で示したものである。



■図5：/ɪ æ ɜː/ のF1/F2 (mel)

学習段階を2つに分けた /æ/ では，2つの段階を比較すると，学習段階1の方が学習段階2よりも，舌の位置が前に位置しており，これが日本人英語学習者にとっての /æ/ の学習であることが分かる。一方，3段階に分かれた /ɜː/ では，学習段階1を基準に考えると，残りの2つのクラスターは舌の高さが学習段階1に近いが舌の位置が後ろよりのものと，舌の高さも低く位置も後ろよりのものと特徴を捉えることができる。前者の方が学習段階1に近いことから，このクラスターを学習段階2，後者を学習段階3と解釈した。/æ/ は初学者が日本語の「ア」に代用してしまう傾向があり，/ɜː/ も「アー」との対比において指導される音である（東後・御園，2009）。従って，/æ/ の学習段階2のクラ

■表9: /i æ ɜː/の学習段階別人数分布と記述統計

	学習段階	*n*	全体発音レベルごとの *n*				平均		SD	
			母語	上位	中位	下位	F1	F2	F1	F2
/i/	1	7	2	4	1	0	447.61	1676.38	31.42	42.42
	2	12	1	1	5	5	448.69	1538.94	34.96	32.10
	3	8	0	3	2	3	539.23	1577.16	25.76	65.37
/æ/	1	12	3	6	2	1	781.60	1380.83	63.09	44.19
	2	15	0	2	6	7	721.53	1244.34	48.76	47.04
/ɜː/	1	13	3	7	1	2	573.43	1328.28	55.28	40.53
	2	4	0	1	2	1	574.81	1176.22	5.23	61.95
	3	10	0	0	5	5	706.54	1148.60	56.29	41.69

Note: F1 とF2の単位はmel である。

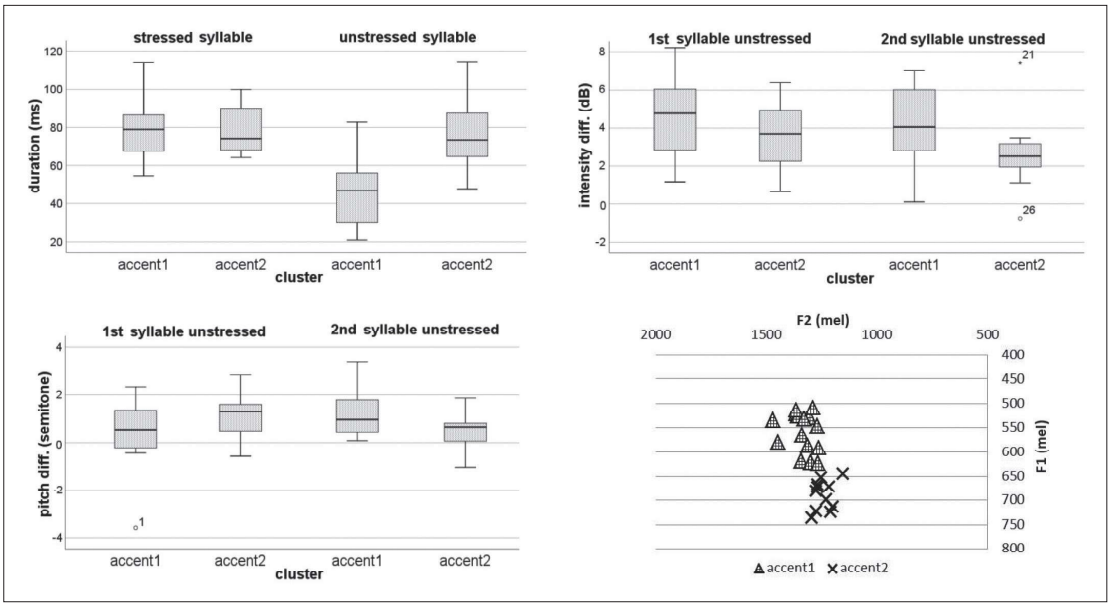
スターと /ɜ:/ の学習段階3のクラスターに重複が見られることは，それぞれがまだ学習の始まりにあるのではないかということを示している。

/ɪ/ の学習段階1は他のクラスターに比べて舌の高さが低いことが分かる。他の2つのクラスターの舌の高さが高めであることから，/ɪ/ の学習とは舌の高さを低くすることであると言えるが，一方で，この2つのクラスターの学習段階の特定はやや難しい。ここでは英語母語話者1人が分類されたクラスターを学習段階2としたが，このクラスターが示す調音位置は，学習段階1の /ɪ/ よりの高さが高く，位置も前にある。これは標準的な英語の /i:/ に近い位置である。それに対して，学習段階3としたクラスターは，学習段階1よりも舌の高さは高いものの舌の位置は近い。学習段階1のクラスターを成す2人の母語話者がアメリカ英語母語話者であること，学習段階2とした1人の英語母語話者がオーストラリア英語話者であることはこうした結果を示した要因のひとつと考えられる。本調査では暫定的に学習段階を上記のように解釈したが，日本語の「イ」や英語の /i:/ のデータと比べ，そもそもの学習の出発点も定義した上で再考する必要はあるだろう。

表9は学習段階別人数と，それぞれの平均値とSDを表したものである。3つの音を比較した場合，/ɪ/ において学習段階1とされた日本人英語学習者が5人と最も少ないため，この音の困難度は高いと言える。清水（1999）のF1とF2の分析においても /ɪ/ は /i:/ とともに日本語の「イ」と重なってしまう音であると報告されているため，本調査の結果もその /ɪ/ の困難さを支持するものであった。清水は /u:/ と /ʊ/ の困難さと，/æ/，/ɑ:/，/ə/ の困難さについても同様に指摘しているものの困難度の違いについては議論していないため，本調査において /ɪ/ が示した結果は新たな示唆を与えるものである。一方で，/æ/ は2つのクラスターに分かれているので，学習段階が2段階となっていることから考えるとこの音が母音の中では最も学習しやすいのではないかとも考えられるが，学習段階が3つに分かれた /ɜ:/ についても，段階は多いものの /æ/ よりも1人多い数の日本人英語学習者が学習段階1とされているので，結果としてどちらがより学習が容易なのかを単純に比較することは難しい。

4.6 アクセント

英語母語話者3人を含んだ学習段階1と学習段階2の2つに分類した。図6は各学習段階について，各変数を視覚的に表した図である。



■図6：強音節と弱音節の長さ（上段左）；弱音節が第一音節，第二音節にある場合の母音との強さの差（上段右）；弱音節が第一音節，第二音節にある場合の母音とのピッチの差（下段左）；弱音節の母音のF1とF2（下段右）

■表10: アクセントの学習段階別人数分布

学習段階	*n*	全体発音レベルごとの *n*			
		母語	上位	中位	下位
1	16	3	8	4	1
2	11	0	0	4	7

■表11: アクセントの記述統計

学習段階	平均								SD							
	強長	弱長	F1	F2	強1	強2	ピッチ1	ピッチ2	強長	弱長	F1	F2	強1	強2	ピッチ1	ピッチ2
1	78.28	45.08	557.80	1336.04	4.38	4.04	0.52	1.24	14.02	17.06	40.22	59.18	2.12	2.20	1.42	0.99
2	79.27	75.67	689.19	1240.71	3.67	2.65	1.13	0.46	13.85	18.03	30.99	42.42	1.95	1.98	0.92	0.75

Note: F1とF2の単位はmel, ピッチ差の単位はsemitoneである。強長 = 強音節の長さ, 弱長 = 弱音節の長さ, 強1 = 弱音節が第一音節のときの後続母音との強さの差, 強2 = 弱音節が第二音節のときの先行母音との強さの差, ピッチ1 = 弱音節が第一音節のときの後続母音とのピッチ差, ピッチ2 = 弱音節が第二音節のときの先行母音とのピッチ差。

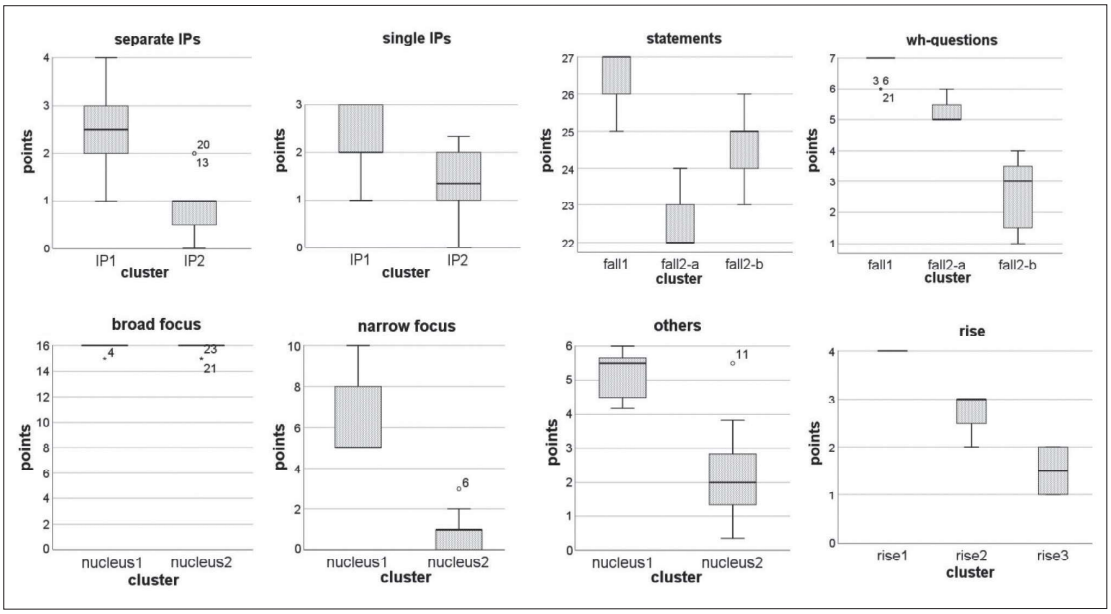
表10と表11は, 学習段階別の人数と, 各変数の平均値とSDをまとめたものである。全体の発音レベルの観点から見ると, 上位の被験者8人が全て学習段階1であったことから, 全体の発音レベルとアクセントの学習には関連があると言える。ただし, 学習段階1と学習段階2の人数分布を見ると, それぞれに分類された人数は13人と11人であり, この結果に基づいて困難度について判断すると, 他の発音項目と比較しても困難度が高い部類とは言えない。一方で, 図6を観察すると, 2つの学習段階の差は主に弱音節の長さと, 弱音節の母音の質にあることが分かる。前者については学習が進むにつれて短くなる傾向があり, 表11の数値を比較すると平均値においては約30 msの差がある。長さに関してはいくつか異なる結果を示す先行研究があるが, 須藤(2010)は機能語の長さについて学習の可能性と困難さの両面を指摘しており, 本調査における結果はそれを支持するものであった。母音の質については学習段階間でさらにはっきりとした境界があり, 図6からも分かるように, 学習段階1の方が舌の高さが高くなり, また位置もやや前寄りとなっていた。波多野・北村(2014)やLee, Guion and Harada(2006)は, 英語圏の滞在経験がある学習者や早期学習者にとってもあいまい母音の生成は困難であることを報告しているため, 本調査の結果はその母音の困難さの裏付けとなった。

4.7 イントネーション

区切り, 音調核の位置, 下降調の核音調, 上昇調の核音調の4項目に分けてクラスター分析を行った。最初の2つの項目については, 3人の英語母語話者を学習段階1として2段階に分類するクラスターとなった。一方で, 下降調は3つのクラスターに分かれたが, 英語母語話者を含む学習段階1の他に形成された2つのクラスターは陳述文の得点が低くwh疑問文の得点が高いクラスターと陳述文の得点が高くwh疑問文の得点が低いクラスターに分かれたため, 学習段階はどちらも2とし, それぞれを段階2-aと2-bとした。上昇調の核音調については, 1人の英語母語話者が他の2人とクラスターが分かれ, かつ最も得点の低い日本人英語学習者2人と同じクラスターへと分類された。しかしながら, その中身を観察すると, この被験者は上昇調の代わりに下降上昇調を使っており, これはまた別のクラスターを形成していた日本人英語学習者と同じ傾向であった。従って, 当該の英語母語話者と同じ傾向を示した日本人英語学習者を他の2人と分かれてしまった1人の英語母語話者と同じクラスターへと修正してから分類して学習段階2とし, 上昇調の箇所で下降調を使う傾向にあった日本人英語学習者2人を学習段階3とした。図7は得点を箱ひげ図によって示したものである。

表12は学習段階別の人数と，各変数の平均値とSDを示したものである。イントネーションの項目間で比較した場合，学習段階1とした人数が最も多かったのは上昇調の核音調の項目であった。Yes-no疑問文ではあまり一般的ではない下降上昇調が英語母語話者1人を含めた被験者に使われたことに関してはもう少し調査が必要ではあるものの，いずれにしても上昇調の学習困難度は低いと言える。そして，上昇調に次いで学習段階1に多くの日本人英語学習者が分類されたのは，区切りに関する項目，下降調の核音調の項目であった。ただし，この結果は全体的な傾向であり，図7が示すように結果を視覚的に見ても，ひとつの項目内であっても変数によって異なる傾向があることが分かる。

第一に，区切りに関する項目であるが，学習段階1となった日本人英語学習者の数は半数を超えたものの，調査文が少ない上に英語母語話者でも全てモデル通りという被験者はいなかった。従って，おそらく困難度は高くないが，もう少し調査を要する項目である。第二に，下降調に関しては，陳述文の得点が低い傾向にあった学習段階2-aの被験者でも，平均値が27点中22.50点と比較的点数が高く，陳述文において下降調を使うべき箇所で8割程度は正しく使うことができていたことになる。一方，wh疑問文に関しては，学習段階2-bに分類された8人の平均値は7点中2.63点と得点が低かった。そのため，下降調の学習困難度も文のタイプによって異なり，全体的に見れば陳述文の方が容易で，ある程度の学習ができていたものと言える。最後に，最も学習段階1の人数が少なかった音調核の位置に関する項目であるが，その中で扱った3つの変数のうち広いフォーカスに関しては学習段階1と学習段階2では大きな差はなく，明らかな違いは狭いフォーカスとその他の文構造などによる場合にあることが分かる。後者の2つに関して，中でも特に狭いフォーカスではほとんど学習が進んでいない。従って，音調核の位置に関する学習では，狭いフォーカスが最も学習の困難度が高く，次いで文構造などによって音調核の位置が決まる文の学習が難しいと言える。



■図7：区切り（上段左2つ）；下降調（上段右2つ）；音調核の位置（下段左3つ）；上昇調（下段右）

■表12: イントネーションの学習段階別人数分布と記述統計

	学習段階	*n*	全体発音レベルごとの *n*				平均			SD		
			母語	上位	中位	下位	IP別	IP1つ	F2	IP別	IP1つ	
区切り	1	16	3	5	4	4	2.56	2.25	–	0.96	0.64	–
	2	11	0	3	4	4	0.91	1.39	–	0.70	0.68	–
							広い	狭い	他	広い	狭い	他
音調核	1	5	3	2	0	0	15.80	6.60	5.17	0.45	2.30	0.79
	2	22	0	6	8	8	15.91	0.86	2.12	0.29	0.79	1.18
							下降陳述	下降wh		下降陳述	下降wh	
下降調	1	15	3	7	4	1	26.53	6.80	–	0.73	0.41	–
	2-a	4	0	1	1	2	22.50	5.25	–	1.00	0.50	–
	2-b	8	0	0	3	5	24.63	2.63	–	0.92	1.19	–
							上昇			上昇		
上昇調	1	22	2	7	7	6	4.00	–	–	0.00	–	–
	2	3	1	1	1	0	2.67	–	–	0.58	–	–
	3	2	0	0	0	2	1.50	–	–	0.71	–	–

Note: IP 別 = 区切りを入れる箇所, IP1 つ = 区切りを入れない箇所。

5 結論

本調査の目的は，日本人英語学習者がELF使用者として学習すべき音声項目について，各項目の学習困難度を音響の側面から検討することである。各項目の音響分析とクラスター分析の結果に基づいた他の音との比較から，清水(2011)に基づいてcoreとした発音項目の中で困難度が高いと考えられるものは，/t/の気音，/v l ɪ/の調音，狭いフォーカスと文構造などに依存して決定される音調核の位置であることが分かった。また，アクセントに関しては全体的に見ると困難度が高いわけではないが，弱音節の短さやあいまいな母音の質を生成することの困難度は高い。一方で，困難度が低いと言えるものは，/w/の調音，上昇調を核音調として使うこと，広いフォーカスの場合に音調核を正しく置くことであった。

こうした結果は，これからの発音指導・学習に向けてELFという視点から示唆を与えるものである。例えば，指導時間や学習時間の観点から言えば，困難度が低いということが分かった項目の指導にはあまり多くの時間を割く必要はないが，困難度が高いと示された項目ついては，時間をかけて繰り返し指導し学習を促す必要がある。また，導入時期に関して考えると，困難度が高い項目ほど配慮を要する。つまり，困難度が高い項目は学習に時間がかかるので早い時期に導入する方が良い，あるいは，困難度が高いから後回しにした方が良いといった単純なことではなく，それらの項目の学習には時間がかかることを考慮しつつ，関連する項目との関係から適切な指導時期を検討する必要があるということである。従って，本調査において困難度が高いという結果が得られた項目については，例えば，気音の中でも/t/は難しいからこそまずは/p/や/k/によって気音の認識を促してから/t/の気音を指導する，/v/, /l/, /ɪ/の調音は，それぞれの対立項目とされている/f b/, /r/, /i: e/の調音指導の後にそれらとの違いを強調しながら指導を行う，狭いフォーカスと文構造などに依存して決定される音調核の位置は，広いフォーカスにおける音調核の位置を指導することで音調核の存在を認識させた後，それらとの対比の中で指

導する方が学習しやすいだろう。

本調査では以上のような示唆が得られたものの，さらなる調査が必要な部分はいくつかある。調査方法の観点から言えば，区切りに関する項目についてはまだ不十分な点があった。また，上昇調や下降調の核音調に関しては，上昇・下降の音声的な側面は捉えていない。ピッチの変動の程度，開始位置といった音声的な観点からの観察を含めてさらに詳細な調査を行うことによって，学習に対するより明確な示唆を与える可能性は十分にある。本調査はその第一歩として位置付けられたものであるということを最後に記しておきたい。

注

(1) 全体の発音レベルを評価する聴覚実験では，本調査用に収集した文章読み上げデータの中から *Getting breakfast* と *Books* (Wells, 2006) を用い，日本人英語学習者を対象として日常的に英語を指導している1人のアメリカ英語母語話者がなまりの程度（accentedness）を9件法で評価し，内容の理解度（intelligibility）を10%刻みの割合で自己評価した。2つの文章の評価の平均値を各被験者のスコアとし，accentedness が3以下，かつ intelligibility が65%以下の被験者を下位，accentedness が7以上，かつ intelligibility が90 %以上を上位と定めた。中位に関しては，上位の被験者が全て英語圏における長期滞在経験を持つ日本人学習者となったため，上位に次ぐ被験者を中位とし，accentedness が5.5から6.5，かつ intelligibility が80%以上の被験者を当該の被験者とした。この聴覚実験には英語母語話者3人のデータも含まれていたが，その結果はいずれも accentedness が9，かつ intelligibility は100%であった。

謝辞

本調査を進めるにあたりご支援くださいました公益財団法人 日本英語検定協会の皆さま，選考委員の先生方には心より感謝申し上げます。とりわけ，調査の道筋に関してご指導、ご助言をいただきました吉田研作先生には深く感謝いたしております。また，調査にご参加いただいた先生方，生徒の皆さんがいらっしゃらなければ，この調査を行うことはできませんでした。快くご協力いただき，ありがとうございました。最後に，貴重な時間を割いて調査全般に渡って相談に乗ってくださった研究室の仲間にも感謝いたします。

参考文献（＊は引用文献）

＊ 足立浩平 (2006)『多変量データ解析法　心理・教育・社会系のための入門』京都：ナカニシヤ出版.

＊ Alptekin, C. (2007). Teaching ELF as a language in its own right: Communication or prescriptivism? *ELT Journal, 61*(3), 267-268.

＊ Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32*(2), 233-250.

＊ Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer (Version 5.2) [Computer program]. Retrieved from http://www.praat.org/

＊ Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer (Version 6.0.05) [Computer program]. Retrieved from http://www.praat.org/

＊ Cogo, A. (2012). English as a Lingua Franca: Concepts, use, and implications. *ELT Journal, 66*(1), 97-105.

＊ Cruttenden, A. (2014). *Gimson's pronunciation of English* (8th ed.). London, UK: Rougledge.

＊ Dewey, M. (2013). The distinctiveness of English as a Lingua Franca. *ELT Journal, 67*(3), 346-349.

＊ 波多野博顕・北村達也 (2014).「日英母語話者による英語弱化母音の音響・調音特徴ーX線マイクロビームデータベースに基づく分析ー」.『日本音響学会誌』70巻3号, 106-113.

＊ Jenkins, J. (2000). *The phonology of English as an international language.* Oxford, UK: Oxford University Press.

＊ Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*(3), 1252-1263.

＊ 北川　彩 (2017).「ELF時代の発音指導に向けて」.『JACET-ICT調査研究特別委員会最終報告書』41-78.

＊ Kuo, I-C. V. (2006). Addressing the issue of teaching English as a lingua franca. *ELT Journal, 60*(3), 213-221.

＊ Kuo, I-C. V. (2007). A response to Cem Alptekin. *ELT Journal, 61*(3), 269-271.

＊ Kent, R. D., & Read, C. (2002). *Acoustic analysis of speech* (2nd ed.). Clifton Park, NY: Delmar.

＊ Ladefoged, P. (2003). *Phonetic data analysis: An introduction to field work and instrumental techniques.* Malden, MA: Blackwell.

＊ Lee, B., Guion, S. G., & Harada, T. (2006). Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals. *Studies in Second Language Acquisition, 28,* 487-513.

＊ Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384-422.

＊ Saito, K., & Lyster, R. (2011). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning, 62*(2), 595-633.

＊ Seidlhofer, B. (2011). *Understanding English as a Lingua Franca.* Oxford, UK: Oxford University Press.

＊ Sewell, A. (2013). English as a lingua franca: Ontology and ideology. *ELT Journal, 67*(1), 3-10.

＊ 清水あつ子 (2011).「国際語としての英語と発音教育」.『音声研究』第15巻第1号, 44-62.

＊ 清水克正 (1999).「英語音声学習における移入ー母音の発音とその音響的特徴」.『名古屋学院大学外国語教育紀要』29, 1-9.

＊ 清水克正 (2008).「L2音声学習とその理論的背景」.『名古屋学院大学論集　言語・文化篇』第19巻第2号, 81-87.

＊ 須藤路子 (2010).「第三章　日本人中学生による英語のリズムパターン習得」.『英語の音声習得における生成と知覚のメカニズムー日本人英語学習者のリズムパターン習得』(pp. 39-57).　東京：風間書房.

＊ Sung, C. C. M. (2013). English as a Lingua Franca and English language teaching: A way forward. *ELT Journal, 67*(3), 350-353.

＊ Sowden, C. (2012). ELF on a mushroom: The overnight growth in English as a Lingua Franca. *ELF Journal, 66*(1), 89-96.

＊ 東後勝明（監修）・御園和夫（編集）(2009).『必携　英語発音指導マニュアル』東京：北星堂書店.

＊ Walker, R. (2010). *Teaching the pronunciation of English as a lingua franca.* Oxford, UK: Oxford University Press.

＊ Wells, J. C. (2006). *English intonation: An introduction.* Cambridge, UK: Cambridge University Press.

第30回 研究助成
C 調査部門 ● 報告Ⅱ ● 英語教育関連の調査・アンケートの実施と分析

発話の諸側面に対する意識の質問紙尺度の開発と妥当性の検証：発話生成モデルの観点から

研究者：英国／ランカスター大学大学院 博士課程　在籍 鈴木 駿吾
《研究助言者：小池 生夫》

概要

本研究では，個人差要因としての「発話に対する意識」を包括的に捉える質問紙尺度の開発と妥当性検証を行う。加えて，本研究の発話意識尺度を用い，発話に対する意識が実際の発話運用に与える影響を考察する。質問紙尺度の作成には，発話生成モデルやCAF指標などの理論的背景に加え，インタビューや発話課題を通した再生刺激法などを用い，学習者の観点を取り入れた。質問紙尺度の妥当性検証には，106名の日本人英語学習者（大学学部生）に対し，質問紙調査を実施し，探索的因子分析を行うことで妥当性の高い質問項目を選定した。その結果，学習者の発話に対する意識は，発話生成モデルの構成要素（概念化，言語形成，調音）を反映したものであることがわかった。加えて，48名の日本人英語学習者を対象に，発話意識尺度と意見を述べる発話課題を実施した。相関分析の結果から，発話に対する意識は，発話運用の発話量，統語的複雑性，正確性，流暢性に影響を与えることが示唆された。

1 はじめに

外国語学習の中でも最も難しいとされるスピーキングには，学習者の言語能力に加え，個人差要因（e.g., 性格，動機付け）や文脈要因（e.g., 話し相手，トピックの複雑さ）などが複雑に関わりあっており，それらの関係性の解明が課題とされている（Segalowitz, 2010）。特に，学習者が発話時に「発話のどの側面に意識を向けやすいか」という傾向は，言語能力と同様に発話に大きな影響を与える。例えば，文法に意識を向けていれば，学習者の発話は文法的に正確になるが，流暢性が阻害され得る。しかし，学習者の意識を包括的に捉えるツールの不足から，実際の発話と発話時の意識の関係性を調査した実証研究は少ない（Vercellotti, 2017）。インタビューを行えば，発話時の意識を詳細に捉え得るが，大規模データでの実際の発話との整合性を調査する際に，そうした質的手法は現実的ではない。そこで，実施が容易で，かつ単一尺度上での比較を可能とする質問紙尺度の開発が重要となる。したがって，本研究では，実際の発話との整合性を外的妥当性として位置づけ，スピーキング時の学習者の発話意識尺度質問紙の開発及び妥当性の検証を行なう。

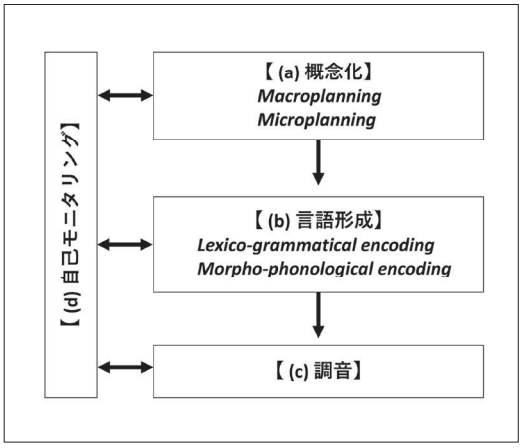
2 研究の背景と目的

2.1 発話生成モデル

まずは第二言語習得理論において，発話生成がどのような言語処理を伴うのかを概観する。これまでいくつかの発話生成モデルが提唱されてきたが（Levelt, 1989, 1999; de Bot, 1992），本研究では，第二言語学習者の特徴である目標言語に関する宣言的知識の重要性を考慮に入れたKormos

(2006)のモデルに基づき論を進める。

伝統的な発話生成モデルと同様に，Kormos (2006)のモデルにおいても発話生成は大きく分けて，(a) 概念化(Conceptualization)，(b) 言語形成(Formulation)，(c) 調音(Articulation) (d) 自己モニタリング(Self-monitoring)の4つの構成要素から成るとしている(図1参照)。発話生成は基本的に(a)→(b)→(c)の順で段階的に行われるが，習熟度が高まると言語知識の自動化により，同時並行的に複数のプロセスを処理することが可能となる。



■図1：発話生成モデル
(Kormos(2006)に基づき，筆者が作成)

最初の(a)概念化の段階では，発話によって達成しようとする目的に応じて，意識的あるいは半無意識的にどのような情報を伝えるべきかを決定し(Macroplanning)，そうした情報の伝達の仕方や順番などを決定する(Microplanning)。この概念化のプロセスによって，言語化すべき概念が特定される(Preverbal message)。次に(b)言語形成の段階では，このPreverbal messageに対応する言語知識を引き出し，目標言語で符号化されたメッセージへと変換することとなる(Internal speech)。言語形成では，語彙，統語，形態素，音韻のレベルで符号化が行われる。それぞれの符号化の順番については議論が分かれるところではあるが，本論ではKormos(2006)に基づき論を進める。まず，概念化で特定された概念が，心的辞書(Mental lexicon)内に蓄積された語彙知識に基づき，レマ(Lemma)へと変換される。一方で，概念同士の関係性(thematic roleなど)が統語的枠組みを決定し，文構造に反映される(節内の目的語などの構成要素や複文などの節間をまたぐ構造なども含む；Agent/動作主→主語)。その後文構造や時制などに応じて，特に動詞や名詞などのレマが定性化(e.g., finite verb)あるいは特定性(e.g., definite article *the*)を形態素として付与され，音韻的に変化する(e.g., takeの過去時制による定性化；take /teɪk/→took /tʊk/)。これらの言語形成を経て符号化されたメッセージ(inner speech)を，調音器官を用いて，空気の振動(音波)として産出する段階が(c)調音となる。

これら(a)-(c)の3つの段階的な処理プロセスに加えて，発話生成中には(d)自己モニタリングが機能している。自己モニタリングは，それぞれの処理段階でのアウトプットが，元となるインプット情報と照らし合わせて適切かどうかを監視し(e.g., 言語形成を経たInner speechが，元のPreverbal messageを適切に伝えることができるか)，問題があった場合に各処理機能にフィードバックを送ることで，言語処理のやり直しを促す。これらの構成要素の関係性を矢印にて図1に示す。

こうした第二言語学習者における発話生成モデルは，モノリンガルの母語話者(以下，母語話者)を対象とした研究を足がかりに発展してきた。そこで，母語話者と第二言語学習者の違いを，上記の発話生成モデルに照らし合わせて再確認したい。第1に，発話生成の構成要素が要する注意の度合い，が挙げられる。母語話者は言語知識が限りなく自動化されているので，(b)言語形成と(c)の調音に関して割かれる注意は第二言語学習者に比べて少ない。一方で，(a)概念化に関しては，言語の習熟度と比較的独立しているので，母語話者・第二言語学習者間で注意リソースを割く量に大きな差はないと考えられている。第2に，第二言語学習者の言語知識量も限られているという点である。例えば，概念化された意味的情報に対応するレマが心的辞書内に存在しない場合は，新たに概念化からやり直すという方略が取られる。また，対応する目標言語での発音情報が習得されていない場合は，L1の発音で代用されるという現象もこの特徴を表す(e.g., 日本人英語学習者の日本語発音)。第3に，第二言語学習者は，多かれ少なかれ目標言語と同時にL1の言語知識も活

性化されるという点である。したがって，活性化されるL1の言語知識を抑制することにも注意リソースが割かれることとなる。

すなわち，母語話者と第二言語学習者の発話の質や速度の違いは，二言語以上の言語知識を持っていることや，不完全な言語知識の自動化・量により，限られた注意リソースで発話生成を行うという点に収束する。したがって，その限られた注意リソースをどのように発話生成中に割り当てているかは，習熟度と同様に発話運用に大きく関わる要因であるといえよう。

2.2 発話の複雑性，正確性，流暢性（CAF）

第二言語習得研究においては，発話能力における習熟度を，実際の発話特徴から捉えようとしてきた。基本的に習熟度（Proficiency）は潜在的な言語知識の集合体と定義され，直接的に観察することは不可能であるとされている（Bulté & Housen, 2014; Lightbown & Spada, 2006）。そこで，実際の言語運用（Performance）やテストなどを通して観察可能な状態にすることで，与えられた課題やテストによって具現化された部分のみであるが，習熟度を部分的に捉えてきた。

発話能力に関しては，先述の発話生成モデルからも示唆されるように，発話能力の多面性を捉えることが最も重要な課題とされてきた（Housen et al., 2012; Michel, 2017）。そうした背景を受けて，Foster and Skehan（1996）による複雑性・正確性・流暢性（CAF）指標の提唱から現在に至るまで，多くの研究者が指標の開発を行ってきた。近年では特にLambert and Kormos（2014）を受けて，理論的妥当性の高いCAF指標を用いることが強く推奨されている。発話能力におけるCAF指標では，先述の発話生成モデルに基づきCAF指標の開発や妥当性検証が行われており，発話にまつわる事象を分析・解釈する上で理論的整合性が非常に高い（cf. Lambert, Kormos, & Minn, 2017; Skehan, 2014）。

また，CAF指標は多面的な発話運用能力を包括的に捉えることができるため，発話運用と動機付けや不安への耐性などの個人差要因との関係性を調査する際にも用いられることが多い（e.g., Dörnyei & Kormos, 2000）。そのような個人差要因の１つとして，発話時に「どのように発話しようとするか」という意識が挙げられるが（e.g., Ortega, 2005; Pang & Skehan, 2014），そうした個人差要因としての意識を包括的に捉えるツールの不足から，実証研究は数が限られている。そこで，発話への意識に関して，先行研究に触れながら，どのような尺度作成が求められているのかを確認したい。

2.3 発話生成における発話時の意識の役割

第二言語学習者の発話は様々な要因の影響を受けることが，第二言語習得研究者の間では周知となっている。近年では，そういった発話運用の複雑性を解釈するために，既存の心理言語学に基づく発話生成モデル（de Bot, 1992; Kormos, 2006）とダイナミックシステム理論（Dynamic System Theory; Larsen-Freeman & Cameron, 2008）を組み合わせた枠組みをSegalowitz（2010）が提唱した。Segalowitz（2010）では，（a）学習者のもつ言語体系（先述の発話生成モデル）に加えて，（b）言語使用に対する学習者の動機付けや態度といった学習者要因（e.g., どのように英語を話したいか），（c）発話を行うコンテクストなどの社会的・外的要因（e.g., 聞き手の人数や準備時間の有無），（d）これまでの発話運用に関する学習経験，といった４つの要因から発話運用を説明することができるとしている。

このSegalowitz（2010）の枠組みでは，比較的安定している平時の発話に対する意識は（b）に該当すると考えられる。一方で，発話課題などの（c）の社会的・外的な要因も学習者の意識に影響を与えるとしている。例えば，限られた時間で情報を伝えなければならない状況では，文法などの正確性よりも流暢性が重視され，意識を払われやすい。つまり，本研究では発話に影響を与える意識というものは，「普段の発話に対する意識」と「発話課題によって変化する意識」の両方を対象とする。また，発話課題そのものが発話者の普段の意識の払い方に影響を与え得るので，普段の発話への意識は，発話課題実施による影響が出ないように測定されるべきだろう。

2.4 発話時の意識を捉える研究手法

発話運用に対する意識の影響を調査した実証研究は少ないが，発話課題実施前の準備時間における認知プロセスや方略などを含む心的努力（Mental effort）について先行研究がなされている。実際の発話中には心的努力と同時に発話生成そのものにも注意リソースが割かれるため，発話課題準備中の方が，より多くの注意リソースが心的努力に割かれることとなる（cf. Skehan, 2014）。したがって，回想インタビューを実施する場合，学習者にとっては課題準備中の心的努力の方が比較的言語化しやすい。これを受けて，先行研究では発話課題準備中の心的努力に焦点を当てて研究がなされた。発話生成モデルに基づけば，発話中と発話準備中では，調音及びそれに関わる自己モニタリングの有無が主な違いとなるが，発話運用における意識の役割を調査する上で重要な知見であることに変わりはない。

まず先駆的な研究としてOrtega（2005）が挙げられる。Ortega（2005）では，発話課題の準備中や実施中の認知プロセスやストラテジーを，学習者の観点から調査するために，44人のスペイン語学習者を対象として，発話課題実施直後に回想インタビューを実施した。結果として，学習者の信念（Learner belief）や志向のような比較的安定した個人差要因としての意識（e.g., 誤用を犯すべきでないと考える）と，発話課題の特徴を受けて反応的に形成される意識（e.g., 適切な単語が思いつかなった場合の対策）の性質の異なる二種類の意識があることを示した。

発話時の意識と実際の発話運用の影響について，Pang and Skehan（2014）では発話生成モデル（Kormos, 2006; Levelt, 1989, 1999）を理論的背景に実証調査を行なった。48人の中国人英語学習者を対象に，絵カード課題による発話の言語特徴（CAF指標）と，回想インタビューによる発話準備中の意識の関係性を調査した。質的データに強く依存していたため，記述的な分析のみであったが，概念化や語彙文法符号化（言語形成）への意識が，発話のCAFに対してそれぞれ協力的・競合的な影響を与える可能性を示唆した。Pang and Skehan（2014）では意識の度合いの強さをインタビュー中の言及回数に頼って操作化を行なったため，相関分析などを実施することができなかった。また，インタビューデータを研究者が発話生成モデルに基づき詳細にコード化を行なったが，そうした理論モデルに基づく区分が，学習者自身の意識の向け方とどこまで整合性があるか調査の余地があると思われる。

これらの先行研究を受けて，発話時の意識に関する質問紙尺度の開発に向けて，考慮すべき点が3つ挙げられる。第1に，「個人差要因としての平時の発話に対する意識」と「特定の発話課題に依存した発話時の意識」それぞれを測定できる質問紙尺度が必要である（cf. Ortega, 2005）。第2に，発話時の意識を連続変数として測定できる質問紙尺度が，発話運用における発話時の意識の役割を調査する上で必要といえる。第3に，質問項目を作成した際には，意識の構成要素の区分が，学習者の観点からみて妥当性の高いものであるかどうかを調査する必要がある。これらを軸に，本研究が行った調査を報告する。

2.5 研究目的

先述の通り，個人の学習者が抱く「どのような発話を行いたいか」という発話への意識は，学習者の習熟度と同様に発話運用に影響を与える重要な個人差要因であるということが叫ばれてきた。しかし，そうした意識を，単一尺度上で，かつ包括的に捉える研究ツールの不足から，回想インタビューをはじめとした質的データに頼った研究が行われており，研究間の比較や一般可能性を考慮した解釈が十分にできない現状が続いている。そこで，発話時の意識と発話運用の関係性を精査するために，本研究は2つの研究目的に沿って，発話に対する意識を測定する質問紙尺度の開発及びその妥当性検証を行う。

第1に，発話時の意識を捉える質問紙尺度（以下、発話意識尺度）の開発と構造的妥当性の検証を行う。発話生成モデルやCAF指標の背景理論に基づき，質問項目の草案を作成する。さらにその草案と併用して，予備調査とインタビューを実施し，学習者の観点から質問項目の修正や追加を行う。加えて，探索的因子分析を行い，作成した発話意識尺度の構造的妥当性を

検証する。

第2に，構造的妥当性の検証を終えた意識尺度を用い，外的妥当性検証として学習者の発話に対する意識が，具体的な発話課題を実施した際に，どのように異なるのかを調査する。さらに，それぞれ「普段の発話への意識」と「特定の発話課題に依存した意識」が，実際の発話運用におけるCAF指標とどのような関連があるのかを調査する。これにより，個人差要因としての発話に対する意識が，実際の発話運用をどの程度説明できるのかを明らかにする。

本論では，第1の研究目的における質問項目作成を調査1，因子分析による構造的妥当性の検証を調査2とした。また，第2の研究目的である発話意識尺度の外的妥当性検証を，調査3と位置付けて本論を進める。

3 調査1：質問項目の作成

3.1 目的

発話生成モデルやCAF指標の背景理論に基づき作成した質問項目を，学習者の観点からみて妥当性の高い項目へと修正及び加筆することを目的とする。また「普段の発話への意識」と「特定の発話課題に依存した意識」の両方に対応できる質問項目を選定するために，それぞれについて探索的にインタビューを行う。

3.2 参加者

インタビューでなるべく妥当性の高い回答をしてもらうために，普段から英語を使用している学生を募集する必要があった。そこで，使用言語が英語である専門科目（English-Medium Instruction [EMI]）を受講している日本人英語学習者（大学学部生）から参加者を募集し，18名が参加した。男性10名，女性8名で，平均年齢は21.72歳（標準偏差 = 0.96）であった。学内で実施されているプレイスメントテストによれば，彼らの習熟度は中級から上級レベルであった。

3.3 手続き

3.3.1 質問項目草案の作成

参加者の学習者としての観点を質問項目に取り入れる前に，本研究の理論的背景としている発話生成モデルとCAF指標に基づき質問項目の草案を作成した。発話生成モデルからは，(a)概念化，(b)言語形成，(c)調音，(d)自己モニタリングを，CAF指標からは(e)文法的複雑性，(f)語彙的複雑性，(g)正確性，(h)流暢性を構成概念の対象として，計12個の質問項目を作成した。参加者には，この質問項目草案に基づきながら，それぞれの項目について想定する具体的な場面や，意識の細部まで口頭で回答してもらった。

3.3.2 半構造化インタビュー

前述の質問項目草案に基づくインタビューの前に，参加者には自由に普段の英語での発話に対する意識について回答してもらうため，半構造インタビューを実施した。Pang and Skehan（2014）を基に以下の３つの質問を作成し，それぞれの質問で根拠や理由を引き出すようにインタビューを展開した。

質問（１）
「普段，英語を話す際にどのように話すよう心がけていますか。あるいは，どんな英語を目標として，普段英語を話されていますか。」

質問（２）
「その心がけや意識は，どのような時に『意識しているな』と実感しますか。」

質問（３）
「そういった意識は，実際に話す時にどの程度役に立っていると感じますか。」

3.3.3 再生刺激法

上記２つのインタビューは「普段の発話への意識」を調査するものである。一方で，「特定の発話課題に依存した意識」を調査するために，再生刺激法（Stimulated recall; Gass & Mackey, 2016）を採用した。回想インタビューでは，参加者自身による考察や意味の付与が加わってしま

い，実際の記憶や認知プロセスとは異なるものを引き出す可能性が高いとされる。そこで，参加者の発話時の認知プロセス（i.e.，発話の諸側面への意識）を，妥当性が高い状態で引き出すことができる再生刺激法は，本調査の目的に合っているといえよう。今回は，参加者の実際の発話を録音したものを刺激とするために，意見を述べる発話課題（Suzuki, 2017; 資料1参照）を用いて，発話を引き出した。再生刺激法では，できる限り記憶が新しいうちに行うことが奨励されているため，発話課題を行った直後にインタビューを実施した。

参加者には，音声再生中に何か思い出した際は，自由に一時停止・巻き戻しをするよう奨励しながら，研究者自身も（a）ポーズの部分と，（b）意味のまとまりの区切り，を目安に音声を停止し，その発話時にどのようなことを考えていたか，なるべくそのままに描写するよう促した。参加者の発話の長さは平均2分45秒（標準偏差 = 1分05秒）であり，インタビューの長さは，平均14分24秒（標準偏差 = 3分16秒）であった。

3.4 分析と結果

上記三種類のインタビューデータを全て書き下ろし，内容分析を行った。今回は従来の質的研究で用いられる純粋な帰納的アプローチというよりも，発話生成モデルやCAF指標に基づき，やや演繹的にグルーピングを行った。

結果，9個のグループに分けられ，各グループから最低3つずつ質問項目を作成した。結果として，41個の質問項目が作成された（資料2参照）。9個のグループのうち5個は，草案と同様に（a）文法的複雑性，（b）語彙的複雑性，（c）正確性，（d）流暢性，（e）自己モニタリングが含まれたが，新たに4つのグループが作られた。概念化を（f）Macroplanningと（g）Microplanningの下位要素に分け（2.1参照），調音を（h）分節（Segmentals）と（i）超分節（Suprasegmentals）の下位要素に分けた。新たに作成された質問項目は複数の大学院生に確認してもらい，文言をわかりやすいものへと修正した。

4 調査2：発話意識尺度の構造的妥当性の検証

4.1 目的

続いて，調査1で発話生成モデルとCAF指標の理論的背景と，インタビューを通して得られた学習者の観点を組み合わせて作成した発話意識尺度の構造的妥当性を検証する。この調査では，「普段の発話への意識」と「特定の発話課題に依存した意識」という二種類の意識の内，前者に焦点を当て，個人差要因としての意識を捉える尺度の妥当性検証を目的とする。

4.2 参加者

調査1と同じ大学にて，日本人英語学習者（学部大学生）男性48名と女性58名，合計106名に質問紙調査を実施した。平均年齢は20.88歳（標準偏差 = 1.57）であった。大学内のプレイスメントテストによると，初級者10名（TOEIC換算で365-555点相当），中級者30名（TOEIC換算で556-689点相当），上級者53名（TOEIC換算で690点以上）であり，106名の内13名はプレイスメントテストの受験経験がなかった。また，参加者の多くは海外経験のない英語学習者であったが，24名が1ヶ月以上の英語圏での滞在を報告した。

調査1と同様，英語での発話に対する意識を調査するため，普段から英語を使用する学生を募集する必要があった。本調査では，比較的大規模な人数が統計手法的に必要なため，EMI受講者に加えて，大学にて英語の授業を受講している学生を対象に募集を行った。つまり，大学入学後も英語学習を継続している学生が本調査に参加した。結果として，比較的習熟度の高い日本人英語学習者が集まったと考えられる。

4.3 質問紙と手続き

調査質問紙には，調査1で作成された41項目の全てを採用した。連続変数とみなすために，“全く意識していない”，“ほとんど意識していない”，“あまり意識していない”，“やや意識してい

る”,“ある程度意識している”,“とても意識している”の6段階尺度を用いた。

質問紙実施の際には，普段発話をする際に達成できているかどうかではなく，どの程度注意を払っているか，について回答することを強調した。また，可能な限り満遍なく選択肢を使うよう指示することで，項目間の意識の度合いを相対的に差別するよう促した。

4.4 分析

各項目の平均値と標準偏差を調査したところ，天井効果や床効果があるものは見つからなかったため，全ての質問項目を用いて，探索的因子分析を行った。Kaiser-Meyer-Olkin measure（KMO）によれば，本調査のデータセットは，質問項目数（観測変数の数）に対して，サンプルサイズが十分であることが示唆された（KMO = .703; cf. Plonsky, 2015）。ただし，いずれの因子にも属さない質問項目が含まれることが予想される。そこで，探索的因子分析を繰り返し，不要な項目を削除しながら，最終因子構造を決定することにした。

4.5 結果

因子抽出法には最尤法を採用し，因子軸の回転にはプロマックス斜交回転を実施した。因子数の決定には，不要な項目を削除しながら，平行分析とスクリープロットを基準に因子数を４因子に決定し，項目数は23個となった（累積因子寄与率：43.9%）。項目削除の基準として，(a) 因子負荷量がどの因子に対しても0.4を下回る項目，(b) 因子に対して2つの項目しか0.4以上の因子負荷量を示さない場合には，該当因子以外に0.4以上の因子負荷量を示さない項目，を削除した。これを繰り返し，最終因子構造とした。モデルの適合度指標であるRMSEAによれば，最終因子構造のモデルはデータに対して当てはまりは悪くはないことが示された（RMSEA = 0.072, 90%信頼区間 = 0.039-0.078）。各項目の因子負荷量と因子間相関を，それぞれ表１と表２に示す。

発話に対する意識に関する先行研究や発話生成モデルの言語処理などの知見を総合し，第１因子を「概念化（Conceptualization）」，第２因子を「言語形成：複雑性（Formulation: Complexity）」，第３因子「言語形成：正確性（Formulation: Accuracy）」，第４因子「調音（Articulation）」に決定した。

5 調査3：発話意識尺度の外的妥当性検証

5.1 目的

Ortega（2005）が示したように，個人の傾向や志向としての「普段の発話への意識」と，「特定の発話課題に依存した意識」の性質の異なる発話への意識を精査する必要がある。また，発話生成モデルやこれまでのCAF研究によれば，学習者が発話のどのような側面を意識するかによって，発話運用に影響を与えることが予測される。そこで，まず「普段の発話への意識」と「特定の発話課題に依存した意識」がどのように異なるのか調査する。その後，発話運用との関連を本研究の発話意識尺度の外的妥当性と位置づけ，意識の度合いが実際の発話運用とどの程度関連があるのかを調査する。

5.2 参加者

調査２の参加者の内，男性20名と女性28名，48人が参加した。平均年齢は20.56歳（標準偏差 = 1.66）であった。大学内のプレイスメントテストによれば，，初級者1名（TOEIC換算で365-555点相当），中級者22名（TOEIC換算で556-689点相当），上級者25名（TOEIC換算で690点以上）であり，1ヶ月以上の英語圏滞在経験がある参加者は11名であった。

5.3 質問紙と発話課題

調査２で普段の発話への意識を測定する際に41項目を用いたため，発話課題時の発話への意識も調査２と同様41項目を使用した。これは参加者に不要な混乱を与えないようにした配慮であったため，分析には調査２の因子分析で特定さ

■表1: 普段の発話への意識尺度項目の因子負荷量

項目番号	質問項目	因子１	因子２	因子３	因子４
19	その時話している内容が、話全体とどのように関連づけられるかを自覚する。	0.860	0.040	-0.052	0.024
20	話全体の構成や情報の伝える順番を考える。	0.708	-0.030	0.017	-0.019
23	話の中で、情報同士の関係が一貫したものになるように心がける。	0.703	-0.108	0.137	-0.051
15	情報や話題が行き来しないように、わかりやすい構成を考える。	0.566	0.207	-0.023	0.086
37	話している最中に、相手に伝えるべき情報の取捨選択をする。	0.519	-0.030	-0.006	-0.014
4	その時話している内容に加え、その先の展開の仕方を考える。	0.448	0.052	0.004	0.220
21	より簡単な語彙で表現する。[a]	-0.286	0.783	0.001	0.150
22	関係代名詞や接続詞などを組み合わせず、なるべく短い文で表現する。[a]	-0.053	0.727	-0.151	0.023
39	自分にとって比較的難しい文法や構文も積極的に使ってみようとする。	0.220	0.570	0.049	-0.019
17	自分が確実に使いこなせる語彙をなるべく用いる。[a]	0.118	0.438	-0.017	-0.189
10	あまり使われない文法や構文も使ってみようとする	0.182	0.425	0.137	-0.097
12	間違いのない英語でなるべく話す。	-0.016	-0.178	0.743	-0.219
38	話の理解を妨げる文法の誤りを起こさないように話す。	0.134	-0.117	0.616	0.068
24	文法的に正しいか自信がなくても、伝わりそうな表現であれば気にせず使う。[a]	-0.098	0.160	0.614	-0.047
13	自分が間違えやすい文法に気をつけながら話す。	0.163	-0.179	0.592	0.002
41	適切な語順で、きちんとした文で話す。	-0.050	0.182	0.582	0.074
25	思いついた語彙が、自分の伝えたい内容とやや異なっていても、伝わりそうであれば気にせず使う。[a]	-0.080	0.127	0.456	-0.112
36	適切な語彙の組み合わせを用いて話す。	0.164	0.132	0.432	0.302
18	区切りに合わせて、リズムを取りながら話す。	0.101	-0.013	-0.183	0.694
35	滑舌よくクリアな発音で話す。	-0.103	-0.060	0.259	0.692
32	単語に応じて、発音の強さ（強弱）を変える。	0.052	0.035	-0.200	0.663
29	なるべく日本語発音にならないように、英語らしい発音を心がける。	-0.326	0.042	0.331	0.556
31	平坦なイントネーションにならないように心がけて話す。	0.182	-0.097	-0.035	0.508
	累積因子寄与率	13.6	11.6	9.9	8.8

N.B. [a] は反転項目をあらわす

■表2: 因子間相関

	因子２	因子３	因子４
因子１	0.262	0.108	0.246
因子２		0.398	-0.052
因子３			0.275

れた23項目のみを用いた。また，調査2と同様に6段階尺度を用いた。発話課題には，調査1と同じ意見を述べる発話課題を用いた（Suzuki, 2017）。本研究の発話意識尺度では，概念化を構成概念として扱うことができるので，絵カード課題のような発話の内容が固定されているものでなく，参加者が内容を考える必要のある内容的に認知負荷の高い発話課題を用いた（資料1参照）。

5.4 手続き

発話課題実施前に，3分間の準備時間を与えた。発話準備中には，ノートの取り方などのテスト方略の要因の影響を最小限にするため，メモなどのノートは取らないように指示した。また流暢性へ過度な意識が向かないように，発話中には制限時間を設けなかった。

参加者は，発話課題実施直後に調査質問紙に回答した。調査2と同様に，発話課題中に達成できていたかどうかではなく，発話課題実施中にどの程度注意を払っていたかについて回答するよう強調し，可能な限り満遍なく選択肢を使うよう指示することで，項目間の意識の度合いを相対的に差別するよう促した。

5.5 分析

5.5.1 意識スコア

調査2の因子構造に基づき，普段の発話への意識と発話課題中の意識において，4つの構成概念それぞれの内的整合性を確認するためクロンバックαを算出した。その結果，発話課題中の意識における「言語形成：正確性」のみ，基準値である.70をやや下回ったが（α = .65），本研究では探索的に意識と発話運用の関係性を捉えることを目的としているため，発話課題中の「言語形成：正確性」に関してもその他の側面と同様に項目間で平均したものを意識スコアとして算出した。

5.5.2 CAF指標

発話時の意識と実際の発話運用の関係性を調査した実証研究は少ないため，以下の3つの基準に基づき，包括的にCAF指標を選出した（Housen et al., 2012; Michel, 2017; Lambert & Kormos, 2014）。

(a) CAFの各（下位）構成概念を理論的に妥当性高く反映できる指標
(b) これまでのCAF研究でよく使われている指標
(c) 他の（下位）構成概念と弁別的な指標

5.5.2.1 発話量（Productivity）

発話量の指標として，総発話語数（Total no. of words）を指標として用いた。直接的にCAFそれぞれの構成概念を反映する指標ではないが，一般的な指標として，発話者が言語化できた情報量の多さを示すため，CAF指標と併用して算出することが推奨されている（cf. Norris & Ortega, 2009）。

5.5.2.2 複雑性（Complexity）

複雑性には統語的複雑性と語彙的複雑性がある（Housen et al., 2012; Michel, 2017）。統語的複雑性指標として，文単位あたりの長さと構造がよく扱われる。本研究ではスピーキングを対象としているので，Analysis of Speech unit（AS-unit; Foster et al., 2000）を文単位とし，(a) 1 AS-unitあたりの平均語数（Mean no. of words per AS-unit）と，(b) 1 AS-unitあたりの平均節数（Mean no. of clauses per AS-unit）を使用する。

語彙的複雑性においては，語彙使用の多面性を考慮し，多様さ（Diversity）・洗練さ（Sophistication）・密度（Density）が主な下位構成概念とされている（Michel, 2017）。それぞれの構成概念を反映するため，本研究では3つの指標を用いた。

多様さには総語数の影響を統制したMTLD（Measure of Textual Lexical Diversity; McCarthy & Jarvis, 2010）を用いた。洗練さには，「低頻出語彙を使用する学習者はより深い語彙知識を持つ」という理論的前提の下，発話内の使用語彙の平均頻出度をCELEX corpusに基づき，CELEX log frequencyを算出した（Graesser et al., 2004）。密度は，テキスト内の意味的情報の割合を指すことから，総語数に対する内容語の割合

(Lexical density)として算出される(Révész et al., 2016)。MTLD と CELEX log frequency は Coh-Metrix 3.0(McNamara et al., 2014)を用い, Lexical density は LexTutor(Cobb, 2011) を用い算出した。

5.5.2.3 正確性(Accuracy)

正確性は, 伝統的に発話中の誤用の頻度として操作化されることが多かったが, 近年では誤用の種類を, 意思伝達を阻害する度合いに基づき重み付けする指標が脚光を浴びている(e.g., 三単現の-s [阻害度低] vs. 誤った語順 [阻害度高])。その中でも妥当性・信頼性共に高い Weighted Clause Ratio (WCR; Foster & Wigglesworth, 2016) を用いた。発話内の各節を4つの水準を用いて評価を行い, 節間の平均値をWCRスコアとした。各水準によって与えられるスコアは, (a) 文法的に正確な節 → 1.0点, (b) 意思伝達を阻害しない誤用を含む節→ 0.8点, (c) 誤用を含むが意図する意味をかろうじて取れる節 → 0.5点, (d) 意思伝達を阻害する誤用を含み, 意図する意味の修復が不可能 → 0.1点, である。

5.5.2.4 流暢性(Fluency)

流暢性もその他の構成概念と同様多面性を持つ。Tavakoli and Skehan (2005) によれば, 流暢性は速度(Speed fluency)・休止(Breakdown fluency)・修正(Repair fluency)3つの側面を持つとされている。日本人英語学習者においても, 流暢性のそれぞれの側面が異なる習熟度の違いを示すことが報告されている(Saito et al., 2018)。加えて, 発話生成モデルの観点からすると, 節内のポーズ(Clause-internal pause) と節間のポーズ(Clause-external pause) では示唆する言語処理が異なるとされている(Götz, 2013; Kormos, 2006)。節内ポーズは発話中の言語形成における非流暢性と関連が高く, 節間ポーズは概念化との関連があるとされている。本研究では, Saito et al.(2018) にならい(a)1分間に産出される平均語数(ポーズ時間を除く)(Articulation rate; Speed fluency), (b) 節外ポーズの割合(Clause-external pause ratio; Breakdown fluency), (c) 節内ポーズの割合(Clause-internal pause ratio ; Breakdown fluency), (d) 非流暢性語彙の割合(Dysfluency ratio)の4つの流暢性指標を用いた。

5.5.2.5 一貫性(Coherence)と結束性(Cohesion)

発話課題を達成する上で, CAF指標のように言語特徴を捉える指標に加えて, どのように意味的情報を表出させているかを示す指標も重要であるとされる。様々な側面がある中で, 発話の一貫性と結束性が重要な手がかりと言える(cf. Kuiken & Veddar, 2017)。本研究では, 一貫性の指標として, 隣り合う文同士の概念的な結びつきの強さを示すLatent semantic analysis(LSA)を用いた。一方で, 結束性においては, 発話内の連結語(Connectives)の多さが結束性の高さを示すとされている。本研究では全体の傾向を探るため, 連結語の種類を考慮せず, 単一指標として1000語あたりの連結語数として標準化したものを算出した(Coh-Metrix 3.0を使用)。

評定者間信頼性を確認するため, 応用言語学での修士号を持つ大学院生1名に, 説明とトレーニングを行い, 発話データのうち25%を用いてAS-unitの数, 節の数, WCRスコアのそれぞれのクロンバックαを算出した。その結果, 全ての項目において信頼性が満たされた(それぞれ, α = .99,.96,.76)。その後の分析には, 発話データを採取した著者の評定を採用した。

5.6 結果

5.6.1. 記述統計量

最終的に算出された意識スコアとCAF指標の記述統計量をそれぞれ表3と表4に示す。意識スコアにおいて, 普段の意識に関しては「調音」が最も高く意識が払われている一方で, 実際の発話課題中の意識に関しては「概念化」について最も高く意識が払われているようである。また「言語形成:複雑性」のみが事実上の最低得点(1.00)を獲得した参加者がいることがわかる。

CAF指標に関しては, 流暢性はあまり高くないが(Articulation rate, 両pause ratio参照), 正確性に関しては比較的高いスコアを示している(0.8 = 意味を阻害する誤用を含まない)。発話能力に関しては, 外国語環境によくみられる初・中

級者を含んだデータセットであると言える(cf. Sato & Lyster, 2012)。

また，外れ値に関して，意識スコアには1名(発話課題中の「言語形成：複雑性」への意識)，CAF指標には様々な項目においてのべ12名の異なるケースが外れ値として検出された。そこで，意識スコアで外れ値となったケースを除外し，CAF指標を伴う統計分析にはノンパラメトリック検定を用いることにした。

■表3: 意識スコア記述統計量

	Cronbach's α	*M*	*SD*	*Min*	*Max*
普段の意識					
概念化	α = .82	3.15	1.00	1.17	5.83
言語形成：複雑性	α = .73	2.01	0.72	1.00	4.00
言語形成：正確性	α = .78	3.15	1.00	1.57	5.14
調音	α = .75	3.74	1.06	1.20	5.80
発話課題特有の意識					
概念化	α = .76	4.08	0.92	2.50	5.83
言語形成：複雑性	α = .65	2.15	0.71	1.00	3.80
言語形成：正確性	α = .71	3.30	0.83	1.29	5.00
調音	α = .82	3.53	1.23	1.40	5.80

■表4: CAF指標記述統計量

構成概念	測定指標	*M*	*SD*	*Min*	*Max*
Productivity	Total no. of words	104.4	47.3	38	250
Syntactic complexity	Mean no. of words per AS-unit	11.61	3.24	6.50	27.78
	Mean no. of clauses per AS-unit	1.55	0.46	0.82	3.56
Lexical complexity	MTLD	51.86	12.67	27.63	89.00
	CELEX log freq.	3.07	0.15	2.62	3.31
	Lexical density	0.51	0.04	0.44	0.60
Accuracy	WCR	0.79	0.10	0.52	0.95
Speed fluency	Articulation rate	112.52	24.50	43.36	168.02
Breakdown fluency	Clause-external pause ratio	0.10	0.03	0.04	0.21
	Clause-internal pause ratio	0.38	0.18	0.14	1.05
	Disfluency ratio	13.69	8.59	0.00	39.68
Coherence	LSA	0.28	0.14	0.01	0.77
Cohesion	Causal connectives freq.	51.91	20.39	0.00	98.59
	Logical connectives freq.	74.36	29.36	20.41	169.01

5.6.2. 発話課題の意識に対する影響

まず，「普段の発話への意識」と「特定の発話課題に依存した意識」では，意識の強さの度合いや払い方が異なるのかを調査する。意識スコアを従属変数として，2つの条件(普段の意識，発話課題依存の意識)と4つの側面(概念化，言語形成：複雑性，言語形成：正確性，調音)を要因として，繰り返しあり二元配置分散分析を行なった。

分析の結果，交互作用が有意であったため（$F(3, 138) = 17.23$, $p < .001$），Holm法を用いて，2条件と4側面間のペアにおける，単純主効果の検定を行なった。その結果，普段の意識において，「調音」がその他3側面と比べて，有意に高く（概念化：$p = .026$, $d = .39$; 言語形成正確性：$p = .026$, $d = .40$; 言語形成複雑性：$p < .001$, $d = 1.25$），「言語形成：複雑性」は 他の3側面と比べ，有意に低かった（調音：$p < .001$, $d = 1.25$; 概念化：$p < .001$, d = 1.11; 言語形成正確性：$p < .001$, $d = 1.19$）。また，「概念化」と「言語形成：正確性」の間には統計的に有意な差は見られなかった（$p = .985$, $d = .003$）。

一方で，発話課題中の意識においては，「概念化」がその他3側面と比べて有意に高く（調音：$p = .027$, $d = .38$; 言語形成正確性：$p < .001$, $d = .60$; 言語形成複雑性：$p < .001$, $d = 1.72$），「言語形成：複雑性」は 他の3側面と比べて有意に低かった（概念化：$p < .001$, $d = 1.72$; 調音：$p < .001$, $d = 0.93$; 言語形成正確性：$p < .001$, $d = 1.11$）。「調音」と「言語形成：正確性」の間には統計的に有意な差は見られなかった（$p = .272$, $d = 0.16$）。

また，条件（普段 vs. 発話課題依存）間での比較では，「概念化」では有意差が見られ（$p < .001$, $d = 0.90$），「調音」において有意傾向が見られた（$p = .052$, $d = 0.29$）。

これらの結果から，発話時の意識は発話課題の影響を受けやすい部分（概念化，調音）と受けにくい部分（言語形成複雑性，言語形成正確性）があることがわかり，発話運用能力や発話生成と同様に発話時の意識にも多面性があることが伺える。次に，そのような発話に対する意識が，実際の発話運用にどの程度影響を与えるのかを調査する。

5.6.3. 発話への意識と発話運用の関連

先述のとおり，CAF指標にはいくつかの指標で異なるケースの外れ値が存在することが確認されたため，CAF指標が関わる分析にはノンパラメトリック検定を用いる。

「普段の発話への意識」と「特定の発話課題に依存した意識」では，意識の払い方が異なることが確認されたので，それぞれの意識スコアと実際の発話課題におけるCAF指標との関係について，スピアマン順位相関係数を用いて検証する。統計的に有意あるいは有意傾向が見られた相関係数とその95%信頼区間を表5に示す。

「概念化」への意識においては，普段の意識の高さが有意に影響する発話の側面は見られなかった。一方で，発話課題中で「概念化」へより強く

■表5: 意識スコアとCAF指標間のスピアマン順位相関係数

意識スコア	普段の意識					発話課題特有の意識				
	CAF指標	*rs*	*p*	95% CI for rs		CAF指標	*rs*	*p*	95% CI for rs	
				Lower	Upper				Lower	Upper
概念化	N.A.					Articulation rate	0.313*	0.032	0.029	0.551
言語形成：複雑性	Total no. of words	0.253†	0.086	-0.037	0.503	Total no. of words	0.310**	0.034	0.026	0.549
	Words per AS-unit	0.275†	0.061	-0.013	0.521	WCR	0.432**	0.002	0.165	0.640
	Connectives freq.	-0.295*	0.044	-0.537	-0.009	Connectives freq.	-0.248†	0.093	-0.499	0.042
言語形成：正確性	Articulation rate	-0.409**	0.004	-0.623	-0.138	N.A.				
	CE pause freq.	0.276†	0.060	-0.012	0.522					
	CI pause freq.	0.246†	0.095	-0.044	0.498					
調音	N.A.					Total no. of words	0.256†	0.082	-0.033	0.506
						Words per AS-unit	0.282†	0.054	-0.005	0.527
						Articulation rate	0.324*	0.026	0.040	0.559

（注）† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

意識することが，より高い流暢性に繋がることが示唆された（Articulation rate）。

「言語形成：複雑性」への意識では，普段の意識の高さが，発話量（Total no. of words）と，統語的複雑性の向上を（Words per AS-unit）促す傾向が見られたが，連結語の使用を抑制する傾向も見られた。一方で，発話課題中での意識は，より強く発話量の増加を促し，同時に正確性の向上（WCR）にも貢献していたが，普段の意識と同様に連結語の使用を抑制する傾向が見られた。

「言語形成：正確性」への意識は，普段から意識をすればするほどに，流暢性を阻害する可能性を示唆したが（Articulation rate，CE pause ratio，CI pause ratio），発話課題中での意識は発話のどの側面にも有意に影響を与えなかった。

「調音」への意識に関しては，普段の意識が実際の発話へ与える影響は見られなかったが，発話課題中の意識が流暢性向上を有意に促し，発話量（Total no. of words）と，統語的複雑性の向上を（Words per AS-unit）促す傾向が見られた。

6 考察

6.1 発話意識尺度の因子構造

発話生成モデルの主な構成要素の内，概念化・言語形成（複雑性，正確性）・調音を反映した4因子に収束した。発話生成には，これらの構成要素それぞれにさらなる下位要素が存在するが，学習者の観点からみた意識では，この4つの構成要素による区分が適切であることが伺える。本研究では，学習者の観点を重視し，意識を「発話者本人が自覚することができるもの」として捉えており（調査1参照），上記の因子構造は妥当性が満たされているといえる。

一方で，自己モニタリングを表現した項目（e.g.，文の途中でうまく表現できないと感じたら，別の言い方を考える）は，いずれの因子に対しても高い因子負荷量を示さなかったため除外された。自己モニタリングは，フィードバックを受けた言語処理の方が自覚しやすいので，自己モニタリングに関わる質問項目が一貫して学習者に理解されなかった可能性がある。例えば，適切な語彙が想起できない場合，言語形成の内の語彙的符号化で問題が生じていることになるので，意識的に概念化から再処理を行うこととなる。このような自己モニタリングが機能した場合では，語彙的符号化の方が学習者本人に自覚されやすいと考えられる。

また，言語形成に関しては，複雑性と正確性という異なる側面を代表する2つの因子に分かれた。似たような意味的情報を言語として表現する際には，微妙な差異はあれども，さまざまな言語方法が存在し得る（i.e., Alternatives; Bloor & Bloor, 2013）。したがって，概念化したメッセージを言語形成する際には，どのような側面を優先して言語形式を選択・採用するかについて，主に2つのスタンダード（i.e., 複雑性 vs. 正確性）が独立して存在するということが考えられる。つまり，これら「言語形成：複雑性」及び「言語形成：正確性」の両構成概念は，(a)概念化したメッセージから言語形成を行う際にどのような言語形式を選択するか，という意識と(b)複雑性と正確性のどちらを優先するか，という個人傾向的な要素を含んだものといえよう。

6.2 個人差要因としての意識

調査2の結果から，発話時の意識には，発話課題の影響を受けやすい側面と，受けにくい側面があることがわかった。今回の研究で用いた発話課題は，認知的負荷の高いものであったため，概念化（e.g.，発話内容の構成など）に対して，優先的に意識が割かれたことは理論的にも説明がつく。発話生成が概念化から始まるように，言語運用は常に内容中心である。したがって，特定の発話課題が概念化にどの程度認知的負荷を与えるかによって，発話課題中の概念化に対する意識の度合いは変容しやすいものであると解釈できる。

また，今回の発話課題では，調音に対する意識が相対的に下がっていた。調音に対する意識の度合いも発話課題によって変容しやすいことが伺える。第二言語学習者の注意リソースは容量が限られており，概念化に意識が割かれたことに

よるトレード・オフ効果である可能性が高い（cf. Skehan, 2014）。

一方で，言語形成に関わる2つの側面（言語形成複雑性・正確性）は発話課題の影響を受けにくいことが示された。このことから，言語形成に関しては，発話課題の特徴よりも，学習者自身の内的要因に依るところが大きいと考えることができる。考えられる要因としては，まさに普段の意識によって定着した個人の傾向が考えられる。あるいは，習熟度や言語知識の量などによって，意識的な言語形式の選択の幅が制約を受けている可能性もある。後続研究が望まれるところである。

6.3 意識が発話運用に与える影響

調査3の相関分析の結果から，発話運用の統語的複雑性，正確性，流暢性の様々な側面に，発話時の意識が影響していることがわかった。以下にそれぞれの意識的側面が発話運用に与えた影響について考察する。

6.3.1 概念化に対する意識

発話課題時の「概念化」への意識が高ければ高いほど，流暢性を向上させた。より意味・内容中心に発話運用を行った場合，流暢性を向上させることが先行研究で言われている（Ellis & Barkhuizen, 2005）。したがって，本研究のこの結果は理論的に説明がつく。一方で，「概念化」への意識は，意味・内容中心の発話運用を意味するが，複雑性・正確性をはじめとする言語形式面を阻害することはなかった。

6.3.2言語形成（複雑性）に対する意識

「言語形成：複雑性」においては，普段の意識と発話課題依存の意識で共通して，発話量を向上させた。発話における発話量は，より多くの情報を言語化していることを示す。つまり，難しい言語表現であっても進んで使おうとする学習者は，言語化することに対しても進んで行うということが伺えるだろう。同様に，「言語形成：複雑性」への意識の高さは，意味的情報を表現する際に高度な言語を使おうとするという意味にも取れる。つまり，内容語に対して，低頻出語を進んで使おうとすると仮定できる。そうした非頻出語は概念的にも難しく，想起するのに多くの注意リソースを費やす。その結果，節と節や，概念と概念との繋がりを明示的に言語化すること（i.e., 結束性）まで意識が届きにくくなり，連結語の使用が抑制されたのかもしれない。

また普段の「言語形成：複雑性」への高い意識は，統語的複雑性（AS-unitの長さ）を促す傾向を示した。多様な表現や比較的難しい言語項目を使うことに対して肯定的であればあるほど，より多くの意味的情報を含めるために形容詞や副詞などの修飾語句を組み合わせ，機能語を用いてより長い文を算出すると考えられる（Norris & Ortega, 2009）。

一方で，発話課題時における複雑性への意識が，正確性向上に貢献することが示唆された。この結果は，意識と発話課題の特徴の相互作用が考えられる。より認知的に難しい発話課題では，高度な言語項目を使用することが促され，完全に習得していない言語項目であっても使わざるを得ない状況になる。そのような状況下では，部分的にしか習得されていない言語特徴を制御できるよう，より明示的な注意リソースが正確性に割かれ，結果として発話運用上でも正確性が向上するとされている（cf. Cognition Hypothesis; Robinson, 2011）。

6.3.3 言語形成（正確性）に対する意識

「言語形成：正確性」においては，普段の意識のみ流暢性を阻害したが，発話の正確性向上には貢献していなかった。これは、発話の流暢性には，言語知識の自動化の度合いだけでなく，発話の正確性に対して払う意識の量が影響することを示しており、流暢性と正確性のトレード・オフ効果といえよう（Ellis & Barkhuizen, 2005）。しかしながら，この結果は，正しく話そうとする正確性への意識が，必ずしも発話の正確性に良い影響を与えるわけではないことを示唆している。つまり，本研究の発話課題においては，発話運用の正確性は意識の影響を受けにくく，言語知識の自動化の度合いが肝要といえる。

一方で，発話課題に特有の「言語形成：正確性」に対する意識は，いずれの発話運用の側面に対して影響を与えなかった。このことからも，内容的

に認知負荷の高い状況では，より限られた注意リソースで言語形成を行う必要があるため，明示的な意識よりも言語知識の自動化の度合いが発話運用の正確性を決定するのではないかと考えられる。

6.3.4 調音に対する意識

「調音」に対する意識に関して，普段の意識の度合いは発話運用のいずれの側面に対しても影響を与えなかった。一方で，発話課題時における「調音」への意識が高いと，流暢性が向上することが示された。「調音」に対する意識が高いことが，流暢性に貢献するのは理論的に整合性がある。具体的には，調音への意識が，発音のリダクション（Reduction）や滑らかなイントネーションを促し，必然的に流暢性が向上したと考えられる。

同様に，発話課題時における「調音」への意識が，発語量と統語的複雑性向上を促す傾向が見られた。この関連は，習熟度による擬似相関の可能性が高い。今回の発話課題のような認知的負荷が高い状況では，内容や意味的情報の伝達が優先される。結果として，内容の構成に関わる概念化や，言語形式の中でも意味的情報に密接に関わりのある語彙や文法の想起に注意リソースが割かれやすい。したがって，認知的負荷が高い状況で，調音まで意識できる学習者は，言語知識の自動化が進んでおり，調音のプロセスにまで注意を割く余裕があると考えられる。つまり，そのような習熟度の高い学習者は，言語化したい意味的情報を捨象することなく言語化できるため，全体の発話量の多さや文単位の長さが高くなる傾向があると考えられる。習熟度という媒介変数が間にあるからこそ，これらの関連は有意傾向に留まったと考えられる。

7 結論と今後の課題

本研究では，発話生成モデルを理論的背景にしながら，学習者の観点を取り入れた発話の諸側面に対する意識の質問紙尺度の開発と妥当性の検証を行った。以下に，本研究で得られた成果を，限界点と合わせて総括する。

まず，本研究においては，発話生成モデルとCAF指標に関する知見を理論的背景とし，質問項目を作成した。加えて，発話への意識に関して，学習者自身の観点が重要視されていることから（Ortega, 2005），参加者に半構造化インタビューや再生刺激法を用いることで，最終的には学習者の観点を取り入れた質問項目へと修正した。

次に，日本人英語学習者を対象に作成された質問項目の調査実施を行い，構造的妥当性検証のために探索的因子分析を行った。結果として，作成した質問項目の内23項目が残り，4つの因子からなる因子構造を示した。それぞれの因子は，「概念化」・「言語形成：複雑性」・「言語形成：正確性」・「調音」と因子名が決定され，学習者の発話に対する意識は，発話生成モデルの主な構成要素の区分と合致することを示した。

上記の構造的妥当性検証に加え，外的妥当性検証として，これまでの発話に対する意識に関する理論や先行研究の成果の再検証を行った。

第1に，「普段の発話への意識」と「特定の発話課題に依存した意識」の性質が異なるかどうかを調査した（cf. Ortega, 2005; Pang & Skehan, 2014）。普段の意識と，意見を述べる発話課題実施時の意識を比較したところ，「概念化」・「調音」は発話課題の影響を受けやすく，「言語形成：複雑性」・「言語形成：正確性」は影響を受けにくいことがわかった。このことから，本研究で作成した質問紙尺度からも，インタビューを用いて検証したOrtega（2005）やPang and Skehan（2014）と同様の結果を支持したということになる。

第2に，これまで記述的な分析でしか行われてこなかった発話時の意識と実際の発話運用の関係性を，量的に精査した。相関分析の結果，普段の意識と発話課題特有の意識とで，発話運用の異なる側面に対して影響を与えることが示された。このことは，本研究で開発した発話意識尺度が測定している発話への意識が，学習者の習熟度やその他の個人差要因と同様に，発話運用に対して何らかの役割を担っていることを意味する。また，発話への意識の役割を調査する際には，普段の意識と発話課題特有の意識それぞれを測定する必要性があるとも言える。

しかしながら，本研究には留意すべき点が少なからず残っている。第1に，本研究の発話意識

尺度の信頼性をさらに調査していく必要がある。本研究では，内的整合性による信頼性のみに頼っているが，今後広く本尺度が使われるためには，再テスト信頼性の確認など，複数の観点から信頼性を検証することが望まれる。第2に，発話課題時の意識では，「言語形成：正確性」に関して，やや内的整合性が下がることが示された。このことから，内的整合性をより高められるように質問項目の文言の修正や追加なども必要であろう。第3に，本研究では発話課題として，意見を述べる発話課題のみを用いて，調査を行った。そこで，発話課題の特徴が発話への意識に与える影響を精査するために，後続研究として，様々な発話課題を用いた調査が望まれる。最後に，本研究は探索的な研究であるため，なるべく包括的にCAF指標の選定を行ったが，発話への意識と発話運用の関連について，研究者が選んだ指標においてのみでしか議論をしていない。本研究で用いたCAF指標では捉えきれていない意識の効果がある可能性を強調したい。

謝辞

本研究を発表する貴重な機会を与えてくださいました公益財団法人 日本英語検定協会と関係者の皆様，ならびに選考委員の先生方に，この場を借りて厚く御礼申し上げます。特に，助言者である小池生夫先生には，本研究に関わることはもちろんのこと，日本や世界の応用言語学などについても非常に的確で，示唆に富むご指導をいただき大変感謝しております。そして，ランカスター大学のJudit Kormos先生，早稲田大学の原田哲男先生には，本研究の理論的意義や実施方法まで多岐に渡ってご助言を頂きました。また，ハワイ大学マノア校博士課程在籍の江口政貴さん，早稲田大学教育学研究科修士課程在籍の杉田歌織さん，同研究科博士課程在籍の安田利典さんの協力なしには，本研究は完遂できませんでした。心より御礼申し上げます。最後になりましたが，本研究にご参加いただいた学生の皆様，誠にありがとうございました。なお，本報告書における不備，遺漏は全て筆者自身にその責があることを申し添えます。

参考文献（＊は引用文献）

＊ Bloor, T., & Bloor, M. (2013). *The functional analysis of English* (3rd ed.). London & New York: Routledge/Taylor and Francis Group.

＊ Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26,* 42-65. http://doi.org/10.1016/j.jslw.2014.09.005

＊ Cobb, T. (2011). The compleat lexical tutor. Retrieved from http://www.lextutor.ca/vp/

＊ De Bot, K. (1992). A bilingual production model: Levelt' s speaking model adapted. *Applied Linguistics, 13,* 1-24.

＊ Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: Psycholinguistic perspective. *Studies in Second Language Acquisition, 20(3),* 349-385. http://doi.org/10.1017/S0272263198003039

＊ Ellis, R., & Barkhuizen, G. (2005). A*nalysing learner language.* Oxford: Oxford University Press.

＊ Foster, P., & Skehan, P. (1996). The Influence of planning and task type on second language performance. Studies in *Second Language Acquisition, 18,* 299-323.

＊ Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language : A Unit for All Reasons. *Applied Linguistics, 21*(3), 354-375.

＊ Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. Annual Review of *Applied Linguistics, 36,* 98-116. http://doi.org/10.1017/S0267190515000082

＊ Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). New York: Routledge.

＊ Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193-202. http://doi.org/10.3758/BF03195564

＊ Götz, S. (2013). *Fluency in native and nonnative English speech.* Amsterdam: John Benjamins.

＊ Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA.* Amsterdam: John Benjamins Publishing Company.

＊ Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, N.J.: Lawrence Erlbaum Associates.

＊ Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321-336. http://doi.org/10.1177/0265532216663991

＊ Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics, 35*(5), 607-614. http://doi.org/10.1093/applin/amu047

＊ Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167-196. http://doi.org/10.1017/S0272263116000085

＊ Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics.* Oxford University Press.

＊ Levelt, W. J. M. (1999). Language production: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 83-122). Oxford: Oxford University Press.

＊ Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, Mass. : MIT Press.

＊ Lightbown, P. M., & Spada, N. (2006). *How languages are learned* (Vol. Oxford han). Oxford: Oxford University Press.

＊ McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-92. http://doi.org/10.3758/BRM.42.2.381

＊ McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

＊ Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition.* (pp. 50-68). Florence: Taylor & Francis.

＊ Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. http://doi.org/10.1093/applin/amp044

＊ Ortega, L. (2005). What do learners plan? Leaner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77-109). Amsterdam: John Benjamins. http://doi.org/10.1075/lllt.11.07ort

＊ Pang, W., & Skehan, P. (2014). Self-reported planning behaviour and second language performance in narrative retelling. In *Processing Perspectives on Task Performance* (pp. 95-127). Amsterdam: John Benjamins.

＊ Plonsky, L. (2015). *Advancing quantitative methods in second language research.* New York, NY: Routledge.

＊ Robinson, P. (2011). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance.* Amsterdam: John Benjamins.

＊ Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37,* 828-848. http://doi.org/10.1093/applin/amu069

＊ Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics, 39*(3), 593-617. http://doi.org/10.1017/S0142716417000571

＊ Sato, M., & Lyster, R. (2012). Peer Interaction and Corrective Feedback for Accuracy and Fluency Development. *Studies in Second Language Acquisition, 34,* 591-626. http://doi.org/10.1017/S0272263112000356

＊ Segalowitz, N. (2010). *Cognitive bases of second language fluency.* London & New York: Routledge.

参考文献（＊は引用文献）

＊ Skehan, P. (2014). *Limited attentional capacity, second language performance, and task-based pedagogy.* (P. Skehan, Ed.). Processing perspectives on task performance. Amsterdam: John Benjamins.

＊ Suzuki, S. (2017). *Investigating Linguistic Dimensions of Second Language Comprehensibility and Perceived Fluency in terms of Complexity, Accuracy, and Fluency.* Unpublished MA dissertation. Lancaster University.

＊ Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.

＊ Vercellotti, M. L. (2017). The Development of Complexity, Accuracy, and Fluency in Second Language Performance: A Longitudinal Study. *Applied Linguistics, 38*(1), 90–111. http://doi.org/10.1093/applin/amv002

資料1：意見を述べる発話課題（Suzuki, 2017）

Argumentative task

Statement:

The Tokyo Olympics in 2020 will bring economic growth to Japan.

How far do you agree?

Give some specific examples, and explain why or why not.

資料2: 調査1に基づき作成された質問項目草案

項目番号	想定した構成概念	質問項目	反転項目
1	流暢性	不自然な場所で英語が止まらないようにする。	
2	統語的複雑性	関係代名詞など多様な接続詞や構文を組み合わせる。	
3	流暢性	速くなりすぎたり遅くなりすぎたりしないように、適切なスピードで話す。	
4	Microplanning（概念化）	その時話している内容に加え、その先の展開の仕方を考える。	
5	自己モニタリング	文の途中でうまく表現できないと感じたら、別の言い方を考える。	
6	Macroplanning（概念化）	準備した内容に加えて、話し続けるために新たなアイデアや具体例などを考える 。	
7	語彙的複雑性	なるべく様々な語彙を用いる。	
8	Macroplanning（概念化）	話している最中は、話し始める前に準備した内容を言語化することに集中する。	✓
9	語彙的複雑性	伝える内容に対して、より厳密な意味を持つ語彙を選ぶ。	
10	統語的複雑性	あまり使われない文法や構文も使ってみようとする。	
11	分節（調音）	相手が一度聞いて何の単語かはっきりわかるように発音する。	
12	正確性	間違いのない英語でなるべく話す。	
13	正確性	自分が間違えやすい文法に気をつけながら話す。	
14	自己モニタリング	自分の発した英語の文法や表現の良否について、話を進めながらも振り返る。	
15	Microplanning（概念化）	情報や話題が行き来しないように、わかりやすい構成を考える。	
16	流暢性	話している最中になるべく止まらないようにする。	
17	語彙的複雑性	自分が確実に使いこなせる語彙をなるべく用いる。	✓
18	超分節（調音）	区切りに合わせて、リズムを取りながら話す。	
19	Microplanning（概念化）	その時話している内容が、話全体とどのように関連づけられるかを自覚する。	
20	Microplanning（概念化）	話全体の構成や情報の伝える順番を考える。	
21	語彙的複雑性	より簡単な語彙で表現する。	✓
22	統語的複雑性	関係代名詞や接続詞などを組み合わせず、なるべく短い文で表現する。	✓
23	Microplanning（概念化）	話の中で、情報同士の関係が一貫したものになるように心がける。	
24	正確性	文法的に正しいか自信がなくても、伝わりそうな表現であれば気にせず使う。	✓
25	語彙的複雑性	思いついた語彙が、自分の伝えたい内容とやや異なっていても、伝わりそうであれば気にせず使う。	✓
26	自己モニタリング	文法的な間違いや伝わりにくい表現に気づいたら、すぐ言い直す。	
27	Macroplanning（概念化）	意見がまとまらないうちに話さないようにする。	✓
28	語彙的複雑性	単語が持つ意味のニュアンスの違いに気をつけて語彙を選ぶ。	
29	分節（調音）	なるべく日本語発音にならないように、英語らしい発音を心がける。	
30	Macroplanning（概念化）	伝えるべき内容に関して、余計な情報を入れすぎないようにする。	
31	超分節（調音）	平坦なイントネーションにならないように心がけて話す。	
32	超分節（調音）	単語に応じて、発音の強さ（強弱）を変える。	
33	自己モニタリング	思いついた表現の意味が通るかどうか、話す前に考える。	
34	超分節（調音）	話の焦点となっている語、大事な部分を強調して発音する。	
35	分節（調音）	滑舌よくクリアな発音で話す。	
36	正確性	適切な語彙の組み合わせを用いて話す。	
37	Macroplanning（概念化）	話している最中に、相手に伝えるべき情報の取捨選択をする。	
38	正確性	話の理解を妨げる文法の誤りを起こさないように話す。	
39	統語的複雑性	自分にとって比較的難しい文法や構文も積極的に使ってみようとする。	
40	流暢性	話が途切れ途切れにならないように一息である程度まとまりのある発話をする。	
41	正確性	適切な語順で、きちんとした文で話す。	

N.B. 項目番号は、質問紙紙面上での順番を示す。