C 調査部門 ● 報告Ⅲ ● 英語教育関連の調査・アンケートの実施と分析

教室における Paired oral test の診断的評価および学習者の受容に関する調査

― 混合研究法を用いて ―

研究代表者：東京都／早稲田大学大学院 在籍 松村 香奈
共同研究者：東京都／早稲田大学大学院 在籍・日本学術振興会特別研究員 守屋 亮
《研究助言者：小池 生夫》

概要 Paired oral testは，面接官と受験者が対話を行う面接型のスピーキングテストとは異なり，英語学習者同士が対話を行うテスト形式である。本研究では，英語を専門としない大学生を対象に，教室でPaired Oral Test（以下，Paired oralとする）を実施した。研究目的は，本研究で用いた評価表の信頼性・教室での実行可能性を検証することと，新形式の英語スピーキングテストおよび受験者の長短所が詳細に提示される診断的フィードバックを学生がどのように受容，認識したかを，コメントシートとインタビューで質的に調査することである。ペアの発話は個別に分析的評価がされ，EBB（Empirically, Binary-choice, Boundary-definition scale）ルーブリック（4.4.1参照）を用い，各観点で得点化した。さらに，各得点から学生の個人属性の推測（プロファイリング）を行うクラスター分析（5.3参照）で探索的にグループ分けのあと，量的データと質的なデータを統合，解釈するという混合研究法を用いた。

1 はじめに

Paired oralは，英語母語話者や英語教員など英語力のある面接官と行う従来の1対1のスピーキングテストとは異なり，自分と同程度の英語力の相手とのより自然な会話を通し，お互いに協力して主体的に対話を維持する努力をする機会を得られるという利点が指摘される（e.g., Galaczi & ffrench, 2011; Taylor & Wigglesworth, 2009）。海外では，英語を母語としない受験者を対象とした，信頼性の高い英語検定試験として，ケンブリッジ英語検定のスピーキングテストがPaired oralの形式で行われており，ヨーロッパ言語共通参照枠（the Common European Framework of Reference for languages: CEFR）（Council of Europe, 2001）にはスピーキングのInteractionの評価基準が示される。近年では，香港，中国本土，韓国といった一部のアジアの国々で大学入試や奨学金制度の資格審査などの試験としてPaired oralが活用されている。一方で，Paired oralを実施する場合，受験者同士の言語習熟度の相違といった言語的な要因のみならず，受験者の年齢，性別，関係性などの要因もテスト結果に影響する可能性があることや（e.g., Brooks, 2009; Van Moere, 2006），評価者の評価の厳しさの相違に関する問題点や（Bonk & Ockey, 2003）妥当性や公平性に問題があると指摘する研究もあるなど（Iwashita, 1996），いくつかの困難が伴う場合がある。このようなPaired oralに特有の難しさから，日本の英語スピーキングテストでは，Paired oralの形式はあまり一般的ではない（Negishi, 2015）。しかし，Koizumi, In'nami, and Fukazawa（2016b）は，（入試や資格試験といった重大なテストではない）教室でのローステークスな（low-stakes）

テストにおいては，教員がさまざまな視点から学生の能力を評価，判断できることから，先に述べたようなPaired oralの困難はさほど問題にならないのではないかと指摘する。また，Swain (2001) は，授業でのペア活動とテストを有機的に結び付けることで，指導と学習におけるポジティブな波及効果があると述べる。

本研究で用いるEBBルーブリックは，学習者のパフォーマンスをもとに実証的に作成するもので(performance-based method)(Fulcher, Davidson, & Kemp, 2011)，作成者によってパフォーマンスサンプルの特徴を弁別して作成され，特定の目的や文脈に応じて用いられる。この作成方法は，専門家の経験や知識から直感的に作成する方法(measurement-driven method)に比較して，各環境でのパフォーマンスの評価に向いているとされ(Fulcher, 1996; Upshur & Turner, 1995; Turner & Upshur, 2002)，今回の研究の文脈に適した評価方法と考える(詳細は4.4.1参照)。

本研究ではPaired oralテストを，学習効果を測る形成的評価とし，教室での授業，診断的フィードバックというアクティビティーの一環として位置づけ，教室内でのテストの実行可能性，ルーブリックの実用性，学生のテストの受容について調査することで，教育的示唆を提供することを目指す。また，クラスター分析による学習者のプロファイリングにより，言語学習支援の視点で調査を進める。

2 先行研究

二者あるいはグループ間での会話に関する研究はテスト形式に限ったものではなく，1970年代にはSacks, Schegloff, and Jefferson (1974) にあるような母語話者間での会話の受け渡し(turn-taking)や話題の展開(topic development)などの談話分析を目的として行われてきた。近年では複数話者での会話のやり取りの研究を目的として，Paired oral あるいはGroup oralスピーキングテストの受験者間の会話の構成を調査・分析が行われている(Galaczi, 2014, p.554)。Kramsch (1986) はinteractional competence (IC)[注1]の用語の紹介と共に，外国語学習指針において，個人の言語習得重視のみで，より良いコミュニケーションの担い手が育成できるわけではなく，話者間の協力姿勢の育成の重要性を述べている。

注1) Young (2011) によれば，相互行為能力(interactional competence: IC)とは，相手との関わりの中で，会話に参加し，個人の言語能力のみならず，非言語資源も用いながら会話を構築する能力を指す。
Communicative competenceが個人の言語スキルに注目するのと異なる概念である。

第二言語(L2)としての英語の習熟度レベルとICの関係については，Galaczi (2014) がケンブリッジ英検のデータを用い，談話分析と会話の受け渡し(turn-taking)やトピックの推移(topic development)に関する談話内での量的分析という手法を用いて，CEFRレベルB1からC2の英語学習者のPaired oralでのICの差異を検証している。IC能力と言語能力は同調して高まり，言語能力の高い学習者はメッセージの産出に長けているという仮説を基に，より良い話者であり良い聴き手となり得ること，そしてそれは，Field (2011) にあるように，作業記憶(ワーキングメモリー)に余裕が生じる結果，より協力的な対話者となり得ると推察している。また，Galacziは，タスクタイプの相違が受験者の習熟度レベルのICにもたらす影響の差異についても明らかにしているが，彼女自身の先行研究と比較して，その影響については比較的小さいと結論付ける。

Iwashita (1996) は，豪州の英語を母語とする日本語学習者(N=20)に対しPaired oralを実施し，6つの観点からなる評価基準を用いて採点し，会話の相手の日本語習熟度レベルが同等の場合と異なる場合を比較した。結果は，相手が自分より高い習熟度の話者の場合にスコアと発話量が高まる場合もあったが，必ずしも全ての受験者に当てはまらないことが示唆された。むしろ，自身の対象言語への不安感や自信の度合い，あるいはタスクタイプがテスト結果に影響すると述べている。また，質問紙の結果から、非母語話者同士

で受験する方が，母語話者が相手の場合より緊張しないことが示された。

日本の英語学習者を対象とした研究では，Negishi（2015）が，大学生24名を対象に5名の評価者で，受験者単独，Paired oral，Group oralの3つの形式でのスピーキングテストを行った。ルーブリックはCEFR-J（ver. 1.1; Tono, 2013）のスピーキングの評価基準を用い，全ての形式と評価者の多相ラッシュモデル適合を確認した上で分析をした結果，Paired oralでやや低い評価が与えられるが，3つの形式での難易度の差は大きなものではなかった。英語習熟度中級レベルの学生（TOEIC©スコア500～700）では，Paired oralやGroup oralにおいて単独よりも高い評価を得ていた。ルーブリックは，レベルによる閾値の逆転の箇所が確認され，更なる検証が必要であるとしている。

Koizumi, In'nami, and Fukazawa（2016a）は，これまで小規模な標本対象の調査が多い中で，日本の大学生167名（内163名を分析）を対象にロールプレーと与えられたトピックでの自由会話で各2種，計4種のタスクを3段階の全体評価（holistic scale）で実施した。一般化可能性理論（第3章参照）での決定研究の結果（3タスクを1評価者）あるいは（1タスクを2評価者）で高い信頼性が得られる可能性が示唆され，多変量解析の一種である構造方程式モデリング（SEM; structural equation modeling）および多相ラッシュモデル（第3章参照）での分析ではPaired oralテストの妥当性が概ね確認された。さらに，同研究チームによりタスクを11に増やし190名の大学生を対象に，先の統計分析手法を用いて信頼性と妥当性の検証が継続された。結果，テストの妥当性，評価表の信頼性が確認された。この研究では，（4タスクを2評価者）あるいは（3タスクを3評価者）で信頼性が十分に確保されたテスト実施の可能性が示された（Koizumi et al., 2016b）。

3 研究課題

本研究は，Paired oralに関する研究が積み重ねられる一方で，英語習熟度レベルが比較的低い学生を対象とした教室でのPaired oralテストについての研究は見られないことから，英語を専門としない日本の大学生を対象に，Paired oralを教室でのアクティビティの一環と位置付け，学習効果を測る形成的評価のためのスピーキングテストとして実施した。教室内でのテストの実行可能性，ルーブリックの実用性，学生のテストの認識について調査することを目的とし，以下の4つの研究課題を設定する。

研究課題1 今回のPaired oralに用いたEBBルーブリックは評価表として適切に機能しているか。

研究課題2 教室での実行可能性の観点から，十分な信頼性を確保するのに必要なタスクと評価者の数はいくつと推定されるか。

研究課題3 Paired oralスピーキングテストのスコアから対象学生をどのようなグループに分類することができるか。また，グループにどのような特徴があるか。

研究課題4 教室でのPaired oralの学生の認識・受け止め方はどのようなものか。

(1) Paired oralを受験後の感想はどのようなものか。

(2) 研究課題3でのグループ間に，認識・受け止めの相違点はあるか。

(3) 従来の面接官・教員と学生という1対1のインタビューテストと比較してどちらの方が難しいと感じるか，またその理由は何か。

研究課題1-3は量的研究で，研究課題4は，量的・質的研究の統合を必要とする混合研究である。研究課題1には多相ラッシュモデル，課題2は一般化可能性理論，課題3はクラスター分析を用い，課題4はテーマ分析を用いた。各分析手法の概略は以下の通りである。

◆多相ラッシュ分析（課題1に対して）

本分析方法では，能力値や難易度の変数を推定

し，データがモデル[注2]にどれくらい適合（fit）しているかを確かめ，評価項目の適切さを吟味することができる。多相ラッシュ分析[注3]では，受験者とタスク以外に，評価者等，3つ以上の相を入れることができる。テストの素点をそのまま利用するのではなく，自然対数を用いてロジット（logit）という単位で数値を求め，その数値をもとに項目の特性や受験者の能力を推定する。これにより，受験者の能力の高さ，観点の難度，評価者の厳しさ，タスクの難度といった詳細な情報を同一の尺度上で解釈することが可能である（図2の変数マップ参照）。

注2）モデルの関数は，受験者変数と項目難易度変数を基に，ロジスティック曲線を用いて導かれる。
注3）分析は，Facetsの簡易版Minifacを用い，観点ごとにレベルの数が異なる場合や観点間で比較的レベルの難易度に差が生じると想定される場合に適したパーシャルクレジットモデル（PCM; Masters, 1982, 2010）で多相ラッシュ分析を行った。

◆多変量一般化可能性理論での分析
（課題1および2に対して）

山森（2004）にあるように，一般化可能性理論とは，分散分析（ANOVA）の原理を用いて，学力やアンケート結果などから得られる心理量の測定において存在する測定誤差の成分と大きさを検討するための方法（Brennan, 2001a）である。テストの測定値に含まれる誤差が，評価者や項目の違いなど何を原因とするものか，また，その誤差の大きさはどの程度なのかを分析することで，十分な信頼性を確保するためには何人の評価者が必要で，いくつのタスクをすればよいかをシミュレーションで検討することができる。いわゆる信頼性係数[注4]にあたる数値については，信頼度指数（Φ-指数：index of dependability）を用いた。実験計画は，複数の評価者が受験者の全てのタスクにおけるパフォーマンスを評価するデザイン[注5]である。分析ソフトは，mGENOVA（Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001b）を用いた。

注4）一般化可能性理論での信頼性係数には「Φ-指数」と「G-係数」の2種類あり，順位付けを目的とする相対評価テストの場合には，一般化可能性係数（G-係数：generalizability coefficient）を用いる。本研究の場合のように，目標とする基準に対する個人の達成度を測る絶対評価テストの場合は信頼度指数（Φ-指数）を用いる（Shavelson & Webb, 1991）。
注5）受験者（p: person），評価者（r: rater），タスク（t: task）3相完全クロス計画（p・ｘ r・ｘ t・）である。

◆クラスター分析（課題3に対して）

クラスター分析は，ある集団の中でデータのパターンが似た人を集めて似た者同士のグループを作る手法である（磯田, 2004）。本研究では，探索的にグループの傾向を探し出す目的で，観点別の得点を基に階層的クラスター分析[注6]を行った。分析はSPSSで実行し，樹形図（図9.デンドログラム）を参考に，最終的には，個々のデータと照らし合わせて確認し，グループ分けが妥当かどうかを判断した。

注6）クラスター化は，個人差研究にふさわしいとされる，似ていない度合を計算する方法として「平方ユークリッド距離」で，また，似ているものを計算する方法として「ウォード法」での分析とした（水本, 2014）。

◆テーマ分析（課題4に対して）

テーマ分析は，収集した各言語データに共通したカテゴリーを，ボトムアップ式にテーマとして抽出する分析手法である（cf., Braun & Clarke, 2006）。本研究では，上記クラスター分析の結果を踏まえ，学習者の英語能力（数量データ）と学習者要因の特長（質的データ）の順序で結びつける，量から質への説明的順次デザイン（Creswel, 2015; 抱井, 2015）で，テーマ分析を行った（e.g., Gkonou, 2018）。今回は特定のフレームワークを用いない探索的・帰納的な分析ではあるが，各データの焦点も明確で，複数のデータを用いたトライアンギュレーション（Denzin & Lincoln, 1994）により，多面的に学習者を捉えることができる。

4 研究方法

4.1 調査協力者

2018年4月から7月にかけて，都内の私立大学に通う日本人英語学習者の協力を得た。対象者は，英語4技能の育成を目的にした必修科目受講者41名である。その内，スピーキングテストを受験した38名をクラスター分析の対象に，振り返りシート2種とインタビューを実施できた34名を質的分析対象とした（授業A,B）。学生の専攻は，人文学，社会学，経営学，日本語学と多岐に渡り，英語専攻の学生は含まれない。英語習熟度レベルはCEFRレベルA2-A1が多く，69名の内訳はIELTS 6.0；1名、英検2級；5名，英検準2級（あるいはTOEIC Bridge130前後）；20名，英検3級；28名，不明；15名である。母語は日本語，海外在住経験3か月以上の学生2名（大学1年次豪州に1年語学留学；1名，英国に入学前に通算3年）。

■表1：参加者、実施文脈および分析の概要

科目名	種別	学生数（男：女）	学年	専攻	分析
授業A	必修	29（13:16）	2年	経営学	（量的）mG-theory，MFRM，クラスター（質的34名）（アンケート）、インタビュー、コメントシート、（フィールドノート）
授業B	必修	9（7:2）	2-4年	様々	
授業C	必修	31（7:24）	2年	社会学	（量的）mG-theory，MFRM

注）mG-theory(multivariate Generalizability theory)は多変量一般化可能性理論、MFRM(many-facet Rasch measurement)多相ラッシュ分析を示す。アンケートおよびフィールドノートはインタビューツールとして用い、分析対象とはしない。

なお，研究課題1および2は，31名（授業C）を含めた計69名で量的分析を行った。授業Cの量的分析の統合にあたっては，評価時期と一部手順が異なったため，授業A・BおよびCを個別に一般化可能性での分析を実施し，受験者変動要因および受験者と評価者の交互作用要因の割合に差がなく（81%:83%，16%:11%），統合に差し支えないと判断した。

4.2 授業および調査実施文脈

授業は一般教室で行われ，初回にオリエンテーションにおいて科目の趣旨説明を研究代表者である担当教員が行った。日本語でペアやグループ活動を足場掛けとして取り入れ（cf.，Gal'perin，1992），その後徐々に学生たちに英語でペアワークをさせることで，英語でのペア，グループ活動にも比較的スムーズに取り組めた。授業は90分で，全体活動60分，個別活動30分の2部構成で，前半は，テキストを用いたコミュニケーション活動を，個別活動では，Paired oralの実施とライティング活動とした。全体活動は，Paired oralの活動に慣れるため，テキストに基づく平易なトピックで学生たちに英語で意見を述べさせるなど英語4技能を統合した活動とした。アクティビティ例としては，学生同士が英語でインタビューをしあう，など実際に英語でのやり取りを必要とするものである。Paired oralは準備が整った第6週目から始まり，授業時間後半30分を使い，複数のペアに研究代表者がPaired oralを，共同研究者が別のPaired oralを終えた学生らにインタビューを行い，他の学生はライティングの課題を進めた。Paired oralとインタビューは後日分析のため共に録音された。

4.3 テスト内容

Paired oralのタスクには，与えられたテーマでの自由会話や（e.g.，Negishi，2015），具体的な場面や文脈で設定役になりきるロールプレー型，二人で話し合って取り組む問題解決型などがある（e.g.，Koizumi，et al. 2016a，b）。しかし，本研究では，与えられたトピックに対して自分の意見を述べ合う議論型（argumentative）タスクを設定した。このタイプのタスクは，立場の表明，理由の陳述，相手への質問，などやり取りが予め半構

造化でき，採点項目が学生と評価者に明確になる。やり取りを半構造化し，指示文に盛り込むことで，言語習熟度レベルの相違による発言の機会の極端な偏りや，沈黙を避ける目的がある。テスト実施者は，英語での会話に不慣れな受験者に，スムーズに会話を促し，テストを実行する方策を施す必要がある（Luoma, 2004）。トピックは2題で，

(1) Which do you think is better, online shopping or instore shopping?（ネットショッピングと店舗での買い物とどちらが良いと思うか）

(2) Some high school students work part-time. Is it good or bad/not good?（高校生のアルバイトは良いか悪いか）

トピックは，テキストにある15個のトピック選択肢の中から，授業でのディスカッションを通し，話しやすいと感じた学生が多かったものの中から最終的に担当教員である研究代表者が選んだ。

4.4 分析・調査用マテリアル

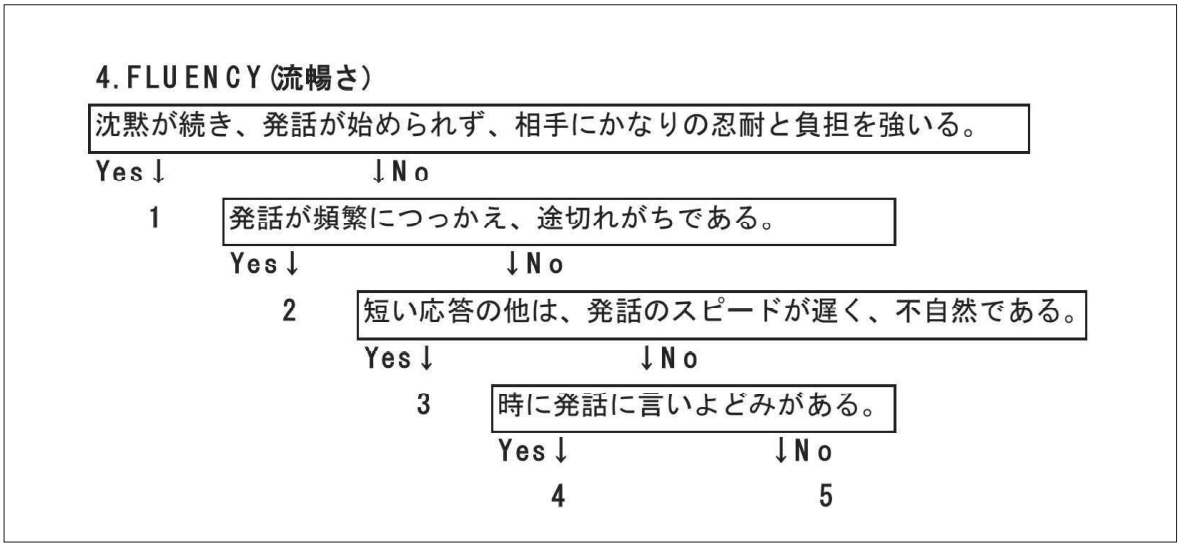
4.4.1. EBBルーブリック

EBBルーブリックは，評価の優先順位の高い順に階層的に配置された記述子（descriptor）[注7]に対し，評価者がYes/Noの2択（binary）を答えていくことで採点結果が得られる（Turner & Upshur, 1995, 2002; Upshur & Turner, 1995）。Knoch（2009）は，こういった作成の経緯の特徴から，診断的評価に適していると述べる。本研究では，図1の流暢さ（Fluency）の他，やりとり（Interaction），内容（Content），および正確さ（Accuracy）の4つの観点でEBBルーブリックを作成した（資料1）。本研究での4観点のEBBルーブリックの内，流暢さ（Fluency）を以下に例示する（図1）。

注7）記述子とは，Yes/Noの判断のための指標となる記述。図1では四角のボックスの中の記述を指す。

4.4.2. 診断的フィードバックシート

Paired oralの結果は，テスト実施の翌週にフィードバックシート（資料2）で学生に提示された。EBBルーブリックをもとに作成されたフィードバックシートは，「達成できたこと」と「取り組むべき課題」が各観点・得点別に一覧でき，評価者は，得点を観点毎に記入し，該当する箇所にマーカーで下線を引くだけで完成する。本形式の利点としては，予めひな型を用意することで，フィードバックの漏れや量のばらつきを防ぎ，記入時間の節約により，迅速に学生への提示が可能となる点である。個別的対応を高めるため，評価表だけでは表現できない内容（各評価に至った具体的根拠等）はシート下部のコメント欄に記入する。学生は自身の得点やパフォーマンスの評価



■図1: 流暢さ（Fluency）観点のルーブリック（数字は得点を示す）

だけでなく，全体の中での位置や，上のレベルで要求される内容も確認できる。久保田（2018）は，受験者に直接EBBルーブリックを提示し，視覚的な効果からも受験者が診断的情報を得られると報告しているが，本研究では，各観点，レベルを一覧にし，達成の可否を対比する形式での提示を試みた。また，事前の英語学習に関するアンケート調査から，英語学習に自信のない学生が多いことから，フィードバックシートはできるだけ肯定的な表現，かつ，学生が次の学習につながるように焦点を絞った文面を心掛けた。なお，本フィードバックシートはインタビュー時の資料として活用された。

■表2: 診断的フィードバック 流暢さ(Fluency)が3点の場合の例

流暢さ Fluency 3／5	◎真摯に伝えようと努力する姿勢が相手に伝わります。	○なかなか話し始めず沈黙すると相手が困ってしまうので、単語レベルでも発してみよう。
	◎適切なタイミングで話し始めることができます。	○途中で言葉が途切れ途切れになったり、沈黙してしまうので気を付けてみましょう。
	◎途中で止まってしまわず、自分なりのペースで話ができます。	○慌てる必要はないですが、聴き手に負担をかけないスピードでの発話を心がけてみよう。
	◎時に詰まってしまいますが、聴き手に負担をかけないペースで話せます。	○速く話す必要はありませんが、相手が安心して聴けるスムーズな発話を目指してみよう。
	◎よどみなく聴き易い発話です。	○口に出す練習を更に重ねていきましょう。

注）◎は「達成できたこ」と、○は「取り組むべき課題」を示す。最上行が１点、最下行が５点に相当する項目

4.4.3. 英語学習アンケートおよび振り返りコメントシート2種

本研究では，調査に先立ち，協力者のこれまでの英語学習状況や背景を把握するために質問紙によるアンケートを行った。アンケートは今回の研究では分析対象とせず，授業の実施やインタビューの際の基礎的な情報として扱った。Paired oral実施後の振り返りシートは次の2種類（資料3）である。

1つ目は，Paired oral実施直後の学生に，2種類のトピックでのパフォーマンスを振り返り，良かった点，悪かった点およびテストの感想を自由記述で記入してもらった。この振り返りシートは，先の診断的フィードバックシートと併せて，インタビュー時の資料として活用した。

2つ目は，授業の最終週に，Paired oralの効果，診断的フィードバックシート感想，および従来の面接官（あるいは教員との）との1対1でのインタビューテストと比較してどちらが難しいか，という以上3点の質問項目に理由と共に自由記述で答えてもらった。

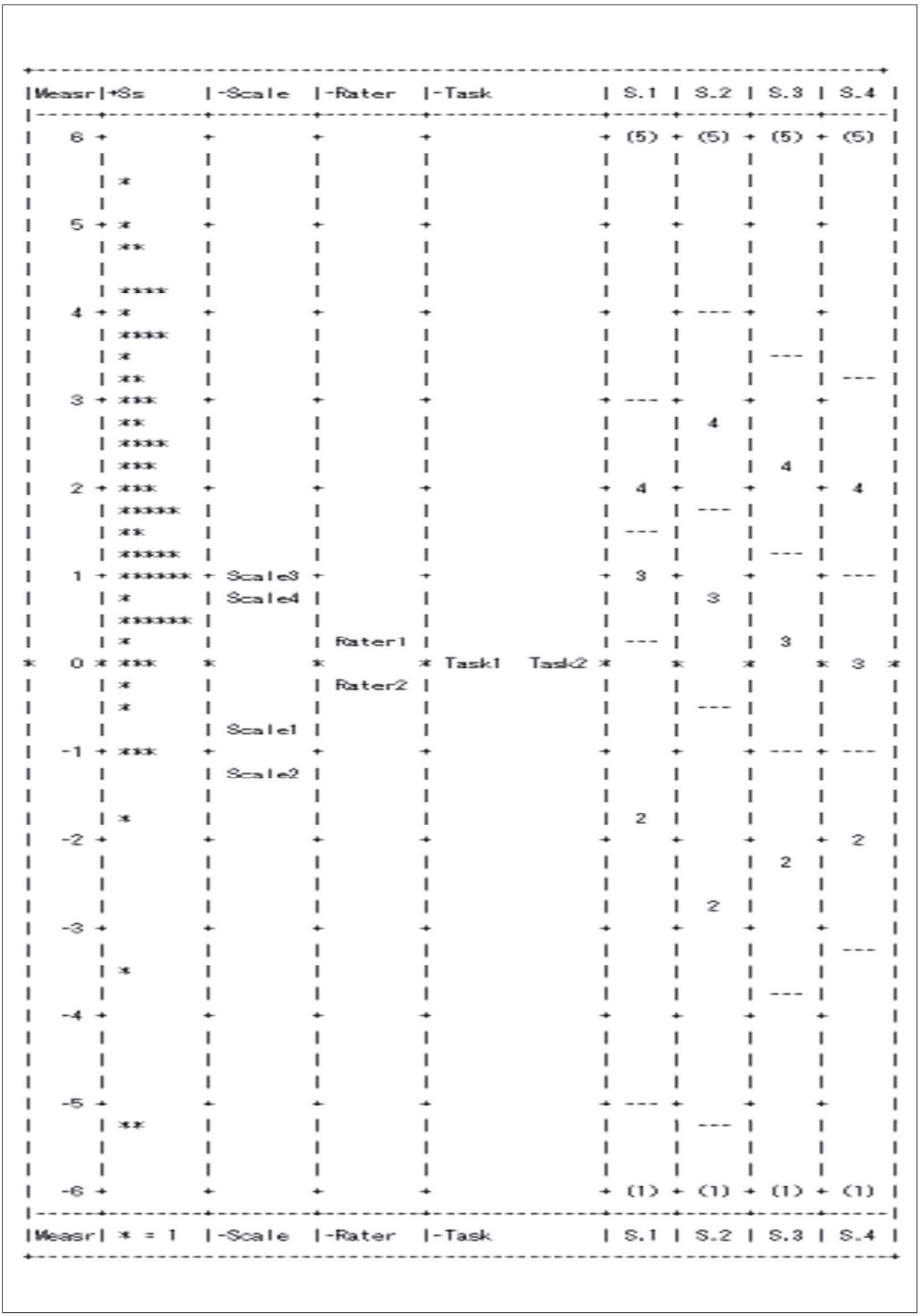
4.5 Paired oral実施手順

トピックは事前に知らせ，実施前までに，教室全体でブレーンストーミング的に考えさせる時間を設けた。ペアの組み合わせはお互いの親しさによる影響を避けるため，実施直前に無作為に教員によって決め，どちらの立場で意見を言うかはテストの場で指示した。教員は，テスト指示書（資料4）を渡し英語で進行役（facilitator）となった。会話は録音され，後日評価された。2つのタスク順による影響が出ないように配慮したが，タスクの難易度に差がないため，タスク順の影響はないとみなした。

5 結論と考察

5.1 研究課題１：EBBルーブリックは評価表として適切に機能しているか

まずは，MFRMの変数マップ（図2）で全体像を確認する。



■図2：能力値・観点難度・評価者厳度・タスク難度の変数マップ

注）左端1列目の数字単位は素点ではなく，自然対数を用いたロジット

各変数マップ(図2)は，左端から，共通尺度で表した推定値ロジット(Measure)，受験者(Ss)，評価観点(Scale)，評価者(Rater)，タスク(Task)の4相を示す。Scaleは，評価観点で，やりとり(Interaction: Scale1)，内容(Content: Scale2)，正確さ(Accuracy: Scale3)，流暢さ(Fluency: Scale4)を表す。右端のS.1-S.4は観点の各Scaleを，垂直方向の1-5は各得点領域を示す。

本図では，テストの素点をそのまま利用するのではなく，自然対数を用いてロジット(logit)という単位で数値を求め，その数値をもとに項目の特性や受験者の能力を推定する。ロジット0を基準に，プラスの値の場合，受験者能力，タスク難度，評価者の厳しさ，タスクの難しさが上がり，マイナスの場合には下がると解釈できる。つまり，上に位置するほど，受験者の能力が高く，評価者は厳しく，タスク，観点は難しいことを示す。このマップでは，Ssはアスタリスク(*)が1人の受験者に相当し，得点による受験者の分布が確認できる。

図2および表3の結果からは，(1)受験者の多くが0ロジット以上に分布し，表3から受験者の平均が1.63ロジットとプラスの値であることから，受験者全体がテストで良い点を取り，取り組み易いテストであったことがわかる。(2)4つの観点では，Accuracy(Scale 3)とFluency (Scale 4)が0ロジット以上で4観点の中ではより難しく，Interaction (Scale 1)とContent(Scale 2)は易しかった。(3)評価者2名の採点の厳しさは同程度であり，共に0ロジットのところに位置しているので，受験者に対して全体を通じて偏った評価をする評価者はいなかった。(4)Interaction (Scale 1)では，得点2の幅が大きく，得点3,4の幅が小さく，レベルが均等に活用されていない。(5)表3のタスクの信頼性が0.00であるのは，平均値，範囲からも，2つのタスクの難易度に差がないことに起因すると推察する(タスクの信頼性については多変量一般可能性理論での成分分析の箇所で更に分析する)。(6)評価観点の分離の大きさは，4つの観点間で難易度の差があり評価のレベルが分散したことを示すと考えられる。(7)受験者の分離の値4.9は受験者が概ね5つのレベルに分けられたことを意味する。

■表3: 4つの相の記述統計 (平均値から範囲までの数字はロジット)

	平均値 (標準偏差)	最小値～最大値	範囲	分離 (separation)	信頼性 (reliability)
受験者	1.63(2.10)	-5.19～5.50	10.69	4.90	.96
評価観点	.00 (1.08)	-1.16～1.0	2.16	11.57	.99
評価者	.00 (.22)	-.16～.16	.32	3.28	.91
タスク	.00 (.01)	-.01～.01	.02	.00	.00

注)一定の歪みは確認されるが正規分布を想定し(基本統計での各相および合計の尖度は-0.7～0.03、歪度は-0.9～-0.3といずれも±2の範囲内)、層(strata)ではなく、分離(separation)での数値を用いる

■表4: モデル適合度統計量の割合(%)

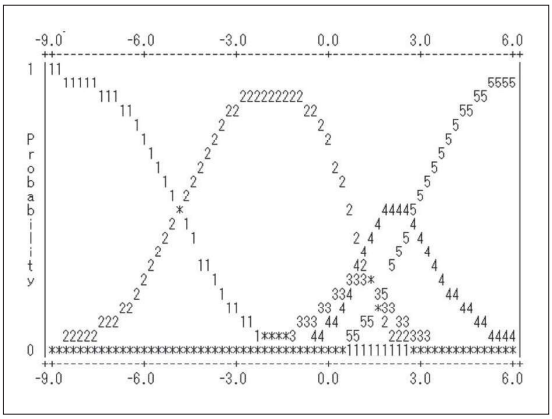
	Overfit	Fit	Underfit
受験者	5.80 (4/69)	89.85 (58/69)	4.35 (3/69)
評価観点	0.00	100.00	0.00
評価者	0.00	0.00	0.00
タスク	0.00	0.00	0.00

注)infit平均平方値が 0.5～1.5(Linacre, 2012)あるいは標準化した値が －2.00～＋2.00(Bond & Fox, 2007)の場合フィットとする

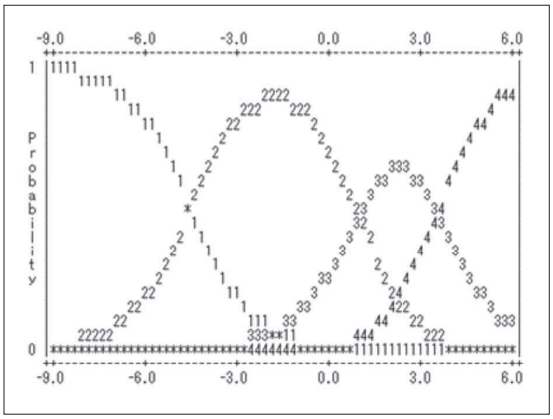
表4は，データがモデルに適合している度合を示す。モデルに実際のデータがどの程度当てはまるか(フィットするか)を分析することで，データの妥当性を確認できる。表4が示すように，受験者項目のみ，7人のデータのみフィットしないことがわかる(Overfit:4; Underfit:3)。この際，問題となるのはアンダーフィット(ミスフィット)の4.35%であり，分析項目から除外するなどの対

応が求められるが,「スピーキングテストなどのパフォーマンステストでは経験上大きめに出る」(小泉, 2016)という報告もあり, テストの目的が診断的フィードバック作成であることから, 分析からミスフィット項目を削除することはしない。

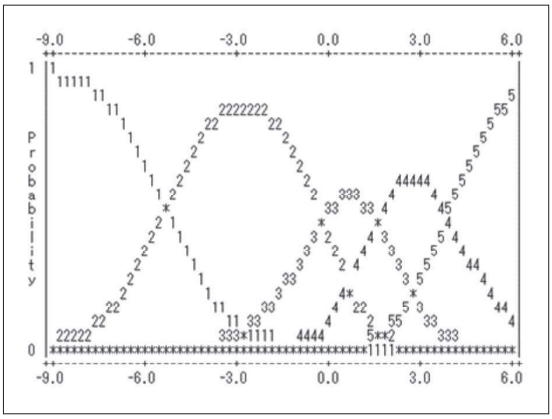
図3-7は,「確率曲線」と呼び, x 軸は受験者能力(ロジット)を y 軸は各得点評価を得る確率を示す。各観点の隣接する閾推定値(表5)が視覚的に確認でき, 評価者の受験者評価で用いた5つの得点領域の形状を把握できることで各観点の得点設定が適切に機能しているか容易に確認できる。



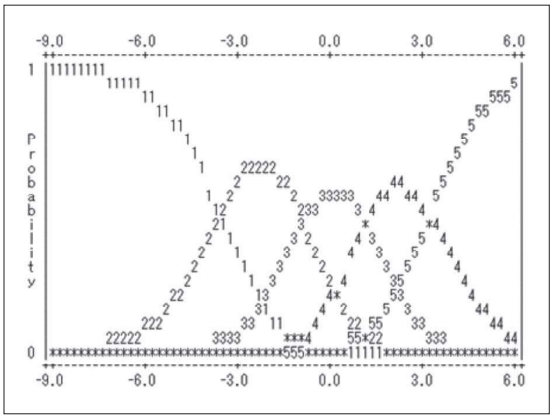
■図3: Interaction (5レベル)



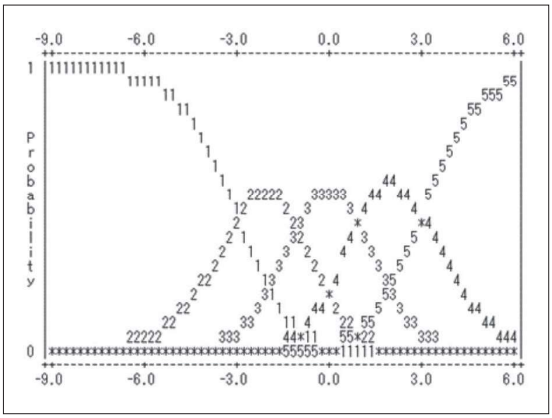
■図4: Tentatively Revised Interaction (4レベル)



■図5: Content (5レベル)



■図6: Accuracy (5レベル)



■図7: Fluency (流暢さ)

評価尺度の適正は，表5の数値とEckes（2015, p.117）の示す次の（a）-（f）の6項目から成る質の高い評価尺度の評価ガイドラインに照らして検証した。（a）各レベルに10以上の得点項目がある。（b）Ratingsの頻度に規則性（各レベルに頂上）がある。（c）各レベルの平均値が一定間隔で増加する。（d）評価尺度のアウトフィット平均平方が2.0以下である。（e）閾が等間隔に増加している。（f）閾の隣同士の距離が1.4ロジット以上5.0ロジット未満である。

結果，Interactionで，複数の項目での逸脱が認められ，得点2と3の閾間の距離が6ロジット，アウトフィット平均平方値が2.0以上あり，得点3と4では閾値の逆転が起きている（図3）。Eckes（2015）の処方箋に則り，試験的にレベルを統合し（得点3と4を3に，得点5を4に変換），4得点レベルにした結果，改善が見られた。今後，EBBの記述子の見直しを含め更に適切な対応を検討したい。その他の観点では，Content観点で，閾間の距離が若干大きいところは確認されたが，Accuracy，Fluencyの観点ともに評価尺度は適切であると判断した（表5）。

■表5: ルーブリックの観点ごとの閾推定値（threshold measures）

	Inter-action	MSu	Revised	MSu	Content	MSu	Accuracy	MSu	Fluency	MSu
レベル1		2.2		2.1		.9		.7		.9
レベル2	-4.88	1.0	-4.53	.9	-5.27	1.2	-3.59	.7	-2.94	.8
レベル3	1.43	1.9	.99	.8	-.21	1.0	-.83	.7	-1.02	.8
レベル4	.76	.6	3.55	1.0	1.58	1.2	1.17	1.1	.94	.9
レベル5	2.69	1.1			3.90	1.3	3.25	.9	3.02	1.2

注）Revisedは、Interactionのレベルを試験的に4レベルにした場合の閾推定値（試験的にレベル3と4を統合し3に、レベル5→4に読み替えた）。
MSuは、アウトフィット平均平方値を表す（2以下であることが望ましい）

多変量一般化可能性理論での4観点の相関は全体に高く，Accuracy-Fluencyが最も高く（0.90），次いで，Accuracy-Content（0.89），Accuracy-Interaction（0.86），一番低かったのがInteraction-Content（0.70）であった。このことからAccuracyが他の観点全体と関りが深く，Interactionが評価観点において比較的独立性が高いことが確認された。

5.2 研究課題2：十分な信頼性を確保するのに必要なタスクと評価者の数はいくつか

次に，教室での実行可能性の観点から，十分な信頼性を確保するのに必要なタスクと評価者の数を推定するために，多変量一般化可能性理論での分析を行った。表6は，得点の分散がどの要因でどの程度説明できるかを示す。

■表6: 決定研究の基準となる観点ごとの推定分散成分（estimated variance components）
（評価者n=2，タスクn=2の場合）

観点 / 変動要因	Interaction	Content	Accuracy	Fluency
受験者（p）	0.89(64.49%)	0.55 (59.14%)	0.73 (63.48%)	0.85 (68.55%)
評価者（r）	0.00 (0.00%)	0.00 (0.00%)	0.04 (3.48%)	0.01 (0.81%)
タスク（t）	0.00 (0.00%)	0.00 (0.00%)	0.00 (0.00%)	0.00 (0.00%)
p x r（交互作用）	0.00 (0.00%)	0.02 (2.15%)	0.03 (2.61%)	0.08 (6.45%)
p x t	0.37 (26.81%)	0.22 (23.66%)	0.19 (16.52%)	0.18 (14.52%)
r x t	0.00 (0.00%)	0.01 (1.08%)	0.00 (0.00%)	0.00 (0.00%)
p x r x t,e	0.12 (8.70%)	0.13 (13.98%)	0.16 (13.91%)	0.12 (9.68%)

まず，テストデザイン（タスクと評価者の組み合わせの数）を推定するための決定研究（D study；Decision study）の基準となる各観点での推定分散成分を確認した。表6ではp：母得点の分散成分（Universe score variance）の割合がいずれも高く，どの観点も分散の60%以上が受験者の能力差で説明できる。つまり，能力を測れていることになる。評価者やタスクの一貫した影響はほとんど見られず，次に高い割合が受験者とタスクの交互作用で，これは，タスクの種類によって，受験者の順位付けが変わることを示す。つまり，学生の中に特定のタスクで取り組みやすさに差がある可能性がある。最後尾の3つの交互作用および残差は，様々な要因が関わっているが，割合が10%程度で問題はなさそうである。前節の表3で，タスクの信頼性が0.0であったが，ある意味，タスクによる大きな影響を受けずに受験者能力が測れるという利点がある。また，表6で受験者とタスクの交互作用が見られることから，難易度に差はなくとも，異なるタスクとしての一定の役割は果たせていると考える。

次の表7は，決定研究により，タスクと評価者の各種組み合わせをまとめたものである。信頼度指数が0.8以上になる評価者とタスクの組み合わせは，教室での文脈を考えると，評価者は1人，あるいはティームティーチングで2人が現実的である。結果，3タスクを評価者2人で，が理想的だが，3タスクであれば評価者が1人でも十分信頼性の高いテストデザインとなる可能性が示唆される。評価者よりタスクを増やす方が有効であることがわかるが，これは表6の受験者とタスクの交互作用での割合から，受験者，評価者にとって取り組みやすいタスクに出会う可能性が高まる効果が期待されるからである。教室実施の実効性とローステイクスなテストの性質から，評価者3，タスク3までの提案とした。

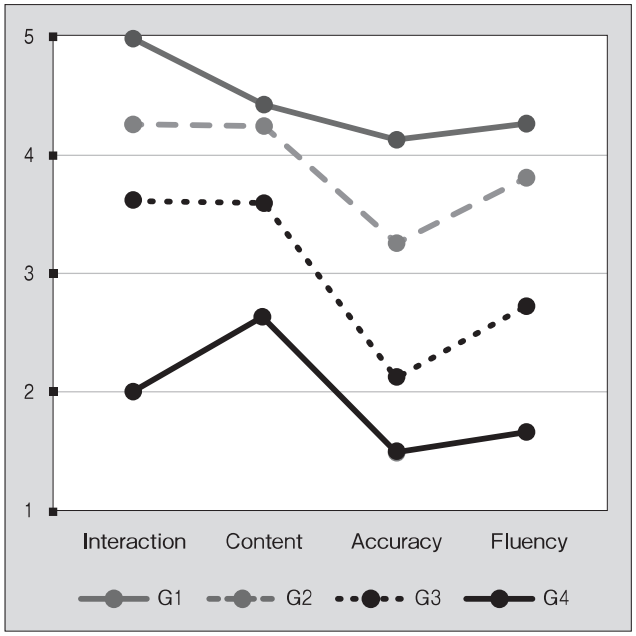
■表7: 決定研究での評価者(r)とタスク(t)の数と観点ごとの信頼度指数(Φ) (pxrxtデザイン)

	r=1 t=1	r=1 t=2	1 3	2 1	2 2	2 3	3 2	3 3
Interaction	0.65	0.78	**0.84**	0.67	**0.81**	**0.86**	0.81	0.87
Content	0.59	0.73	**0.80**	0.65	0.78	**0.84**	0.80	0.85
Accuracy	0.63	0.74	0.79	0.70	**0.80**	**0.85**	0.83	0.87
Fluency	0.68	0.78	**0.82**	0.75	**0.84**	**0.87**	0.86	0.89

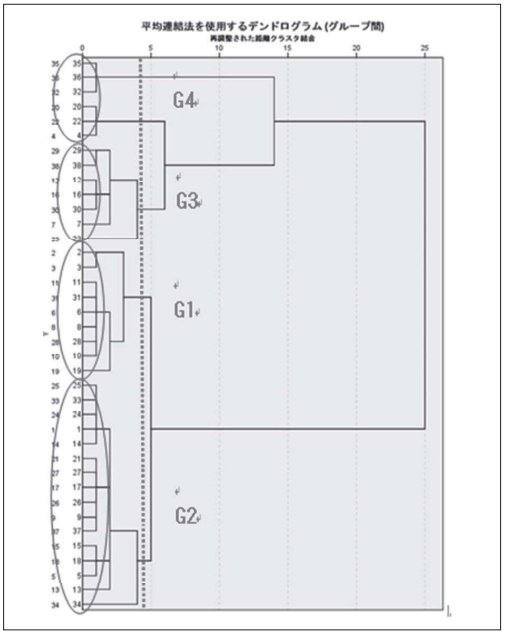
注）下線はΦ指数が0.8以上の項目。太字は、推奨の組み合わせとその信頼度指数

5.3 研究課題3：Paired oralのスコアから対象学生をどのように分類できるか

1つの試みとして，Paired oralのスコアから対象学生を，G1，G2，G3，G4の4グループに分類することができた（図8）。クラスター分析の分類のための変数として，4つの観点の各スコア（2人の評価者の平均得点）および合計点を用いた。合計点を用いることで，結果的に習熟度レベル別に分かれることになったが，分類は受験者のフィードバック目的ではなく，教室での全体の指導を念頭に，レベル別の特徴を探索した。なお，同じ大学での今回とは異なる学生群を対象にしたパイロット研究でも，4グループに分類でき，本研究対象の学校の教室での学生はPaired oralの視点では，概ねこのような4レベルのグループに分類できると推測できる。図9および表8が示すように，Contentは予め内容を準備できるため，観点の中では差が小さく，AccuracyとFluencyは英語力の違いが現れるのか，レベルごとに均等に差が出ている。デンドログラムでの第一分岐はG1・G2グループとG3・G4グループだが，これはこの2つの観点に見られる能力差に起因すると思われる。特徴的なのはG1のInteractionの高さとG4のInteractionの低さである。総合的な英語力の高さは，余裕をもったやり取りにつながる可能性が高い。反対に，やり取りに必要な聞き取りの能力，語彙や文法を基礎とする英語産出力のいわゆる「言語」の部分での問題点が，半構造化した比較的取り組みやすいタスクであっても，G4には難しかった可能性がある。グループごとの相違については，次節の課題4（2）で質的データと組み合わせて分析する。



■図8：4グループの４つの観点での平均値比較



■図9：デンドログラム（樹形図）

■表8: グループごとの各観点の平均値と標準偏差

グループ	人数	Interaction	Content	Accuracy	Fluency
G1	7	5.0 (0.0)	4.4 (0.5)	4.1 (0.4)	4.3 (0.5)
G2	17	4.3 (0.7)	4.3 (0.6)	3.3 (0.5)	3.8 (0.6)
G3	8	3.6 (0.5)	3.6 (0.9)	2.1 (0.6)	2.8 (0.7)
G4	6	2.0 (0.0)	2.7 (0.8)	1.5 (0.6)	1.7 (0.5)
Total	38	3.9(1.1)	3.9(0.9)	2.9(1.0)	3.3(1.1)

5.4 研究課題４

5.4.1 研究課題４（１）：Paired oralを受験後の感想はどのようなものか

テスト受験直後の振り返りシート（感想）と学期末でのPaired oralに対する認識（効果）を要約したものに実際のコメントの中から典型的だと思われるものの抜粋を記載した（表9）。

「感想」は，「肯定的」，「否定的」および「改善点」に分類できた。全体的に「否定的」内容が多いが，この結果を英語学習に関するアンケート[注8]の学習者背景から推察すると，英語学習に対する自己評価が他の構成概念と比べて低い[注8]ことがその一因かもしれない。教室での授業活動の一環のテストにも関わらず，とても高い緊張感を述べる学生が多かった。「効果」は，どの観点において効果を感じたかに注目し，主に3種類に分類された。その中の，「言語・非言語」は英語の4技能や発音・語彙等に関するコメントや相手とのインタラクションの際に欠かせないジェスチャーや頷き，アイコンタクト等に関しての内容である。「自律学習」に関しては，主に学習者の心理的要因（動機付け，自信，感情等）や学習方略に関するものが含まれる。全体的にはPaired oralに直接関係するような「言語・非言語」といった内容が過半数を占めたが，その副産物として心理的要因や学習方略といった自律学習に関する内容を多くの学生が述べていたのは注目に値する。

注8）学習者アンケートは，自己調整学習（3.4）・目的（4.3）・自己評価（2.7）・モチベーション（3.6）・充実感（3.0）・自己受容（3.6）の6項目計37問から成る質問紙を6件法で回答してもらったものである。（　）内の数字は平均。各項目のαは，.75-.87であった。

診断的フィードバックについては,「今までの成績は, なぜSABCDが付いたかわからなかったが, 自分の良かったところ, 直すべきところがわかり, 評価がしっかりしていてうれしい」「何を頑張って何から始めればいいかが具体的に書いてある。先生からの直接の評価は新鮮」,「評価項目が細かく, 1人1人へのコメントもあり, 自分の良いところ悪いところが明確」「決してネガティブでない表現で書かれていて優しさを感じた」など, 自分の弱点や強みを把握し, 英語学習につながる影響が確認できる。

■表9: Paired oral に対する学生の受け止め(数字はコメント数)

コメント	内訳	詳細(抜粋)
感想(65)	肯定的(14)	「自分の体験がある話題で考えて答えられてよかった」,「相手の意見を最後まで理解できたのはよかったと思う」,「偶然意見が固まっている方の役割だったためよかったけれど, 反対側の意見を聞いてなるほどと思った」
	否定的(30)	「実際に喋ると自分の単語力の無さを感じた」,「とても緊張した」,「もっと練習してれば詳しく言えたと思い, ちょっと後悔しました」
	改善点(21)	「自分の意見に理由だけでなくもう1つくらい追加で話せたらよかったと思う」,「単語やフレーズなどしっかり覚えていかなければならないと感じた」,「質問された内容に正しく答えられるようになりたいです」
効果(110)	言語・非言語(65)	「相手の意見を正確に聞き取る能力」,「リアクションの方法が解った」,「英語を話す良い機会だった」「暗記でなく相手の反応を見ながら会話できた」「英会話の基本がわかった」「発音の大切さ認識」「相手への伝え方を理解」
	自律学習(41)	「留学にさらに行きたくなった」,「具体的に何を勉強すべきか理解」,「英語に対するハードルが下がった」
	その他(4)	「新しい視点を発見」,「新しい価値観を得た」,「話したことのない人との交流」

5.4.2 研究課題4(2):グループ間に、認識・受け止めの相違点はあるか

次に, Paired oral直後の振り返りシートの記述を分析対象とし, 良かった点, うまくいかなかった点を4グループで比較した。

表10は, テストで良かった点, 達成できた点についての感想で,「内容」と「言語・非言語」の意見に大別された。各グループの人数が最小6名(i.e., G4)から最大17名(i.e., G2)と大きく異なるため, 慎重に考察する必要があるが, グループが上がるにつれ「内容」に関しての比率が大きくなることが明らかになった。G1とG4を比較してみるとより明白で, G1とG4の「内容」と「言語・非言語」を見てみると, その割合はほぼ逆転している。この結果は先行研究を踏まえても, 学習者の英語能力が上昇するほどタスクの際の認知的負荷が減り, 認知的リソースを他の高次認知機能に費やせるようになるという社会文化理論や認知心理学等の知見とも一致する(e.g., Lantolf, Poehner, & Swain, 2018; Ortega, 2009)。

前述の傾向は, 表11にも見られる。G1の方が「言語」に対する割合が若干高いが, これは先述の表10とも照らし合わせてみると「内容」や「非言語」が満足にできたことの裏返しであり,「心理」的緊張等が少なく,「言語」の点を挙げ, 認知的リソースに余裕がある分, 細部に目を向けられたと考える。

■表10: Paired oral受験直後の感想①（良かった点・達成できた点）

グループ	内訳	詳細（抜粋）
G1（14）	内容（12）	「その場で思いついた事を，事前に考えていたこととあわせて伝えることができた」，「考えてきたことを英語で表現することができた」
	言語・非言語（2）	「相手の目を見て話せた。緊張したけど最後まで文章を言えた」，「相手の目を見て話しができた」
G2（31）	内容（16）	「自分が言いたかったことを大体言えた点」，「相手の主張を汲み取る事を気をつけた」
	言語・非言語（15）	「文法ミスは少なかったと思います」，「相手の反応を見て相づち等のレスポンスができた方だと感じた」
G3（14）	内容（7）	「言いたいことをなんとか時間はかかったが言えた」，「相手の言いたい事は理解でき，それに答えられた」
	言語・非言語（7）	「難しめの文を作れた」，「何度も英語で話したり練習したりしていたので英語を使えた」
G4（11）	内容（2）	「どうにか伝えられた」，「良い点をうまく伝えられた。自分なりに」
	言語・非言語（9）	「初めの決まり文句を覚えておいたおかげで極度の緊張はなかった」，「相手の意見を聞くときは相手のことを見ることができた」

注）括弧内はコメント数を示す。学生数は、N=38,G1（n=7），G2（n=17），G3（n=8），G4（n=6）

■表11: Paired oral受験直後の感想②（悪かった点・うまくいかなかった点）

グループ	内訳	詳細（抜粋）
G1（14）	言語（9）	「単語が出てこなくて，出てきた単語を話してしまった」，「英文をスムーズに話せなかった」
	非言語（2）	「指示文を見過ぎて相手の目を見ていなかった」，「ジェスチャーやアイコンタクトが不足していた」
	内容（3）	「もっと追加で話せばよかった」，「もうちょっと詳しいことが言えればよかった」
G2（30）	言語（15）	「正しいかどうかわからなかった。流暢さに欠けた」，「言いたかったことを即座に言えなかった。文法がめちゃくちゃだった」
	非言語（2）	「相づちがめんどくさくてやらなかったこと。頷くくらい」，「途中、相づちを忘れた点」
	心理（6）	「頭が真っ白になって言葉がでなかったこと」，「緊張して上手く喋れなかった」
	内容（7）	「理由と具体的な内容があまり一致していないと思った」，「本当の意見は?と聞かれた時に，general reason（?）しか答えられずに終わってしまった…」
G3（15）	言語（13）	「正しい文法に自信が持てなくて迷った」，「全く頭が働かなくて日本語が出てしまった。思いがけない質問に止まってしまった」
	心理（2）	「緊張がすごかった」，「途中、緊張があり出なくなった」
G4（9）	言語（9）	「何を話せばいいのか浮かばなかった。英単語が全く出てこなかった」，「自分で考えていた文章通りに言うことができなかった」

注）括弧内はコメント数を示す。学生数は、N=38,G1（n=7），G2（n=17），G3（n=8），G4（n=6）

5.4.3 研究課題４(3):1対1のテストと比較してどちらの方が難しいと感じるか

面接官や教員と1対1のインタビュー形式のテストとPaired oralのどちらが難しいかを尋ねた結果は表12の通りである。82%(38人中31人)の学生は,英検や高校時代の英会話の授業などインタビュー形式のテストの経験はあったが,Paired oralの経験は全員が初めてであった。回答はこれまでの経験との比較で,各人により異なるため,あくまでも参考データとしての扱いである。全体ではPaired oralの方が難しいと感じる学生(53%)が,1対1の方が難しいと感じる学生(31%)より多かったが,グループ別で内訳をみると,中級レベルのG2だけがPaired の方が難しいと感じる割合が高かったが,G1,G3およびG4では,難しさが変わらないと感じていたのは非常に興味深い。G1の学生は「どちらとも言えない」という学生を含め,意見が3分され,G4は両形式で2分するが,両グループとも,どちらの形式も「言語」を測るテストとして捉えているのに対し,G2では,よりPaired形式の「相手の存在」や「即興性」を意識しているようである。

6 結論

混合研究法の視点から,Paired oralに関する各観点の得点である量的結果を4グループに分けたグラフと,Paired oralに対する学習者の受け止めに関する質的データの要約を統合し,ジョイントディスプレーで示したのが図10である。全ての観点で高得点のG1では,Paired oralを言語テストの一形態として適切に捉え,強化すべき点を認識し,英語学習に結び付けられる。G2では,自分と同等の他者との会話という強い意識を持ち,相手への配慮の必要性と同時に,英語上級者との会話とは異なる難しさを認識しているようであった。G3,G4では,自分の英語力の不足,特に,聞き取りや語彙力の不足から,発話が思うようにいかないことを自覚しつつ,簡単な言い回しなどでコミュニケーションが取れ,自分の意見を英語で述べることのできる達成感を感じたようであった。

最後に,量的・質的それぞれの分析に加えて混

■表12: Paired oralと1対1のインタビュー形式のテストのどちらが難しいか

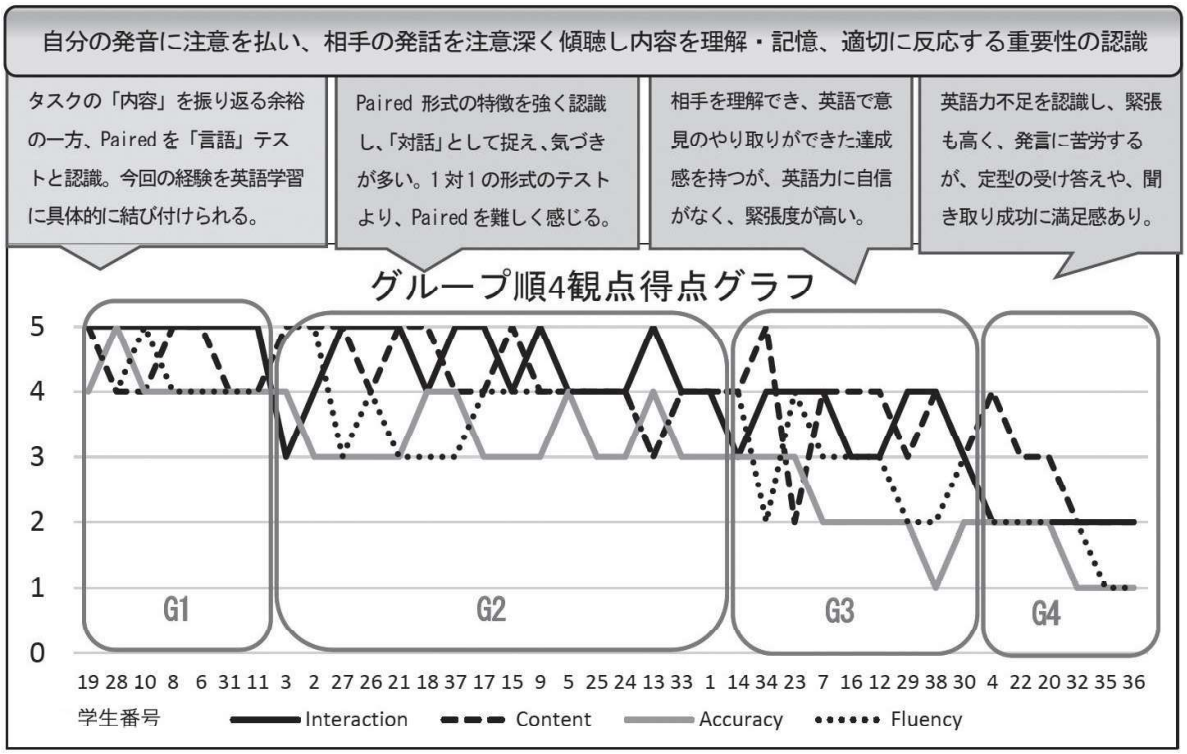
	Pairedの方が難しい	どちらとも言えない	1対1の方が難しい
G1(n=7)	2	3	2
G2(n=17)	11	2	4
G3(n=8)	4	1	3
G4(n=6)	3	0	3
Total(N=38)	20(53%)	6(16%)	12(31%)
理由 ( )は 4グループを示す	●「アドリブや英語力が必要(G2)」 ●「先生は言いたいことを読み取ってくれるが、生徒の方は理解できるようにと気を遣う。相手の言いたいことも理解する努力が必要(G2)」 ●「相手の話の主旨を聞き取り、相手に聞き取りやすい発音で返答するという実質的な英語力が必要(G4)」	●「(Pairedは)相手によって聞き取りにくいかも、でも先生の英語は上手すぎて聞き取りにくいから(G1)」 ●「どちらにしても知らない相手で緊張する(G1)」 ●「結局文法力と単語力がないと話せないのは同じ(G2)」	●「Pairedは普段の授業でのペア活動のように意見交換ができる(G1)」 ●「Pairedの方が対等に会話。先生だとこちらが合わせる感じになる(G2)」 ●「英検では先生が本当に何を言っているか全然わからなかった。生徒だと少し聞き取れて予測できる(G2)」 ●「緊張して何も話せなくなる(G4)」

合研究法により4つの研究課題から明らかになったことをまとめる。RQ1:作成されたEBBルーブリックの妥当性，信頼性については，やり取り，内容，正確さ，流暢さの4つの観点のうち，やりとりの観点は記述子の再検討やスケール修正が必要であるが，全体として信頼性の確保された評価方法であると推定できた。RQ2:決定研究（D study）の結果から，半構造化され，評価基準が明確なargumentativeなタスクであるという条件の下であるが，1人の評価者でも3つのタスクがあれば，信頼性が確保され，教室でのテストとしての実行可能性が示唆された。RQ3およびRQ4:教室での学習活動の一環として行った彼らの初めてのPaired oralは，評価者による診断的フィードバックやテストをツールとしてのインタビューの機会により，自らの英語対話力を客観的に認識，受容する機会となったと考えられる。全体の受け止めとして，英語での対話には，特に自分の発音に注意を払い，相手を注意深く傾聴し，発話内容を理解・記憶，適切に反応する重要性を強く感じること，そして何より，他者と英語で意見交換ができたという達成感があることが確認された。一方，Paired oralテストの結果から，学生全体は概ね4つのグループに分類することができ，同じ教室内でもグループによってPaired oralに対してそれぞれ特徴的な認識があり，相応の言語支援が望ましいと考えられる。

今後の課題は，教室の外にも目を向け，Paired oralが，英語熟達度を反映した他のテスト得点と関係があるかどうかを測る外挿（Extrapolation）の視点での妥当性の検証を進める必要があると考える。

謝辞

本研究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，また，本稿の執筆にあたり，ご助言，丁寧なご指導を賜りました小池生夫先生には深く感謝申し上げます。早稲田大学の澤木泰代先生には，ゼミでの研究指導に対し，心より感謝申し上げます。今後とも研究に精進してまいりたいと思います。



■図10: Paired oral 個人総合得点順4観点の量的結果およびテスト受容記述のジョイントディスプレー

参考文献（＊は引用文献）

＊ Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

＊ Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology,* 3(2), 77-101.

＊ Brennan, R. L. (2001a). *Generalizability theory,* New York: Springer.

＊ Brennan, R. L. (2001b). Manual for mGENOVA. Version 2.01. Iowa: Iowa Testing Programs, University of Iowa.

＊ Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366.

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

＊ Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: John Wiley, (Available from Books on Demand, University Microfilms, 300N. Zeeb Rd., Ann Arbor, MI 48106)

＊ Creswell, J.W. (2015). *A Concise Introduction to Mixed Methods Research,* Thousand Oaks, CA: Sage Publications.

＊ Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research.* Thousand Oaks, CA: Sage publications.

＊ Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* (2nd ed.). Frankfurt am Main, Germany: Peter Lang.

＊ Field, J. (2011). Cognitive validity. In L. Taylor (Ed.): *Examining speaking: Research and practice in assessing second language speaking.* (pp. 65-111).Cambridge University Press.

＊ Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing,28*(1), 5-29. doi:10.1177/0226553220935931 4

＊ Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking test? *Applied Linguistics, 35,* 553-574. doi: 10.1093/applin/amt017.

＊ Galaczi, E., & ffrench, A. (2011). Context validity. In L. Taylor (3d.), *Examining speaking: Research and practice in assessing second language speaking* (pp.112-170). Cambridge, UK: Cambridge University Press.

＊ Gal'perin, P. Y. (1992) Stage-by-stage formation as a method of psychological investigation. *Journal of Russian & East European Psychology, 30*(4), 60-80.

＊ Gkonou, C. (2018). Listening to highly anxious EFL learners through the use narrative: Metacognitive and affective strategies for learner self-regulation. In R. L. Oxford & C. M. Amerstorfer (Eds.), *Language learning strategies and individual learner characteristics: Situating strategy use in diverse contexts* (pp. 79-98). London, UK: Bloomsbury Academic.

＊ 磯田貴道 (2004).「英会話テストの信頼性の検討ー一般化可能性理論ー」三浦省五（監修）前田啓明・山森光陽（編著）磯田貴道・廣森友人（著）『英語教師のための教育データ分析入門 授業が変わるテスト・評価・研究』(pp. 112-124).東京：大修館書店

＊ Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5,* 51-66.

＊ 抱井尚子(2015).『混合研究法入門 質と量による統合のアート』,東京：医学書院

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

＊ 小泉利恵(2016).Facetsを使った多相ラッシュ分析ーパフォーマンステストの妥当性検証に向けて．早稲田大学,言語教育とデータ分析に関する連続ワークショップ(2016/12/23)資料

＊ Koizumi, R., In' nami, Y., & Fukazawa, M. (2016a). Development of a paired oral test for Japanese university students. In C. Saida, Y. Hoshino, & J. Dunlea (Eds.), *British Council New Directions in Language Assessment: JASELE Journal Special Edition* (pp. 103-121).

＊ Koizumi, R., In' nami, Y., & Fukazawa, M. (2016b). Multifaceted Rasch analysis of paired oral tasks for Japanese learners of English. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 89-106). Gateway East, Singapore: Springer Singapore. doi:10.1007/978-981-10-1687-5

＊ Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern language Journal, 70*(4), 366-372.

＊ 久保田恵佑 (2018). 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-. *STEP Bulletin, 30, 42-66.*

＊ Lantolf, J. P., Poehner, M., & Swain, M. (Eds.) (2018). *The Routledge handbook of sociocultural theory and second language development.* New York, NY: Routledge.

＊ Luoma, S. (2004). *Assessing Speaking.* Cambridge: Cambridge University Press.

＊ Masters, G.N. (1982). A Rasch model for partical credit scoring, *Psychometrika, 47,* 149-174.

＊ Masters, G.N. (2010). The partial credit model. In M.L.Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp.109-122). New York, NY: Routledge.

＊ 水本篤 (2014).「多変量解析入門 不空数の変数を同時に分析するには」竹内理・水本篤（編著）『外国語教育研究ハンドブックー研究手法のより良い理解のためにー』(pp.181-193). 松柏社

＊ Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *Annual Review of English Language Education in Japan, 26,* 333-348.

＊ Ortega, L. (2009). *Understanding second language*

参考文献（＊は引用文献）

acquisition. New York, NY: Routledge.

＊ Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, Vlo.50,* No.4, Part1 (Dec., 1974), pp.696-753.

＊ Shevelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A primer.* Newbury Park: Sage Publications.

＊ Swain, M. (2001). Examining diaglogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18* (3), 275-302.

＊ Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26,* 325-339. doi: 10 1177/0265532209104665

＊ Tono, Y. (2013). *CAN-DO list sakusei katsuyo, eigo totatsu-do shihyo* [A CEFR-J guidebook. CAN-DO list development and utilization; Reference for English achievement]. Tokyo: Taishukan.

＊ 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

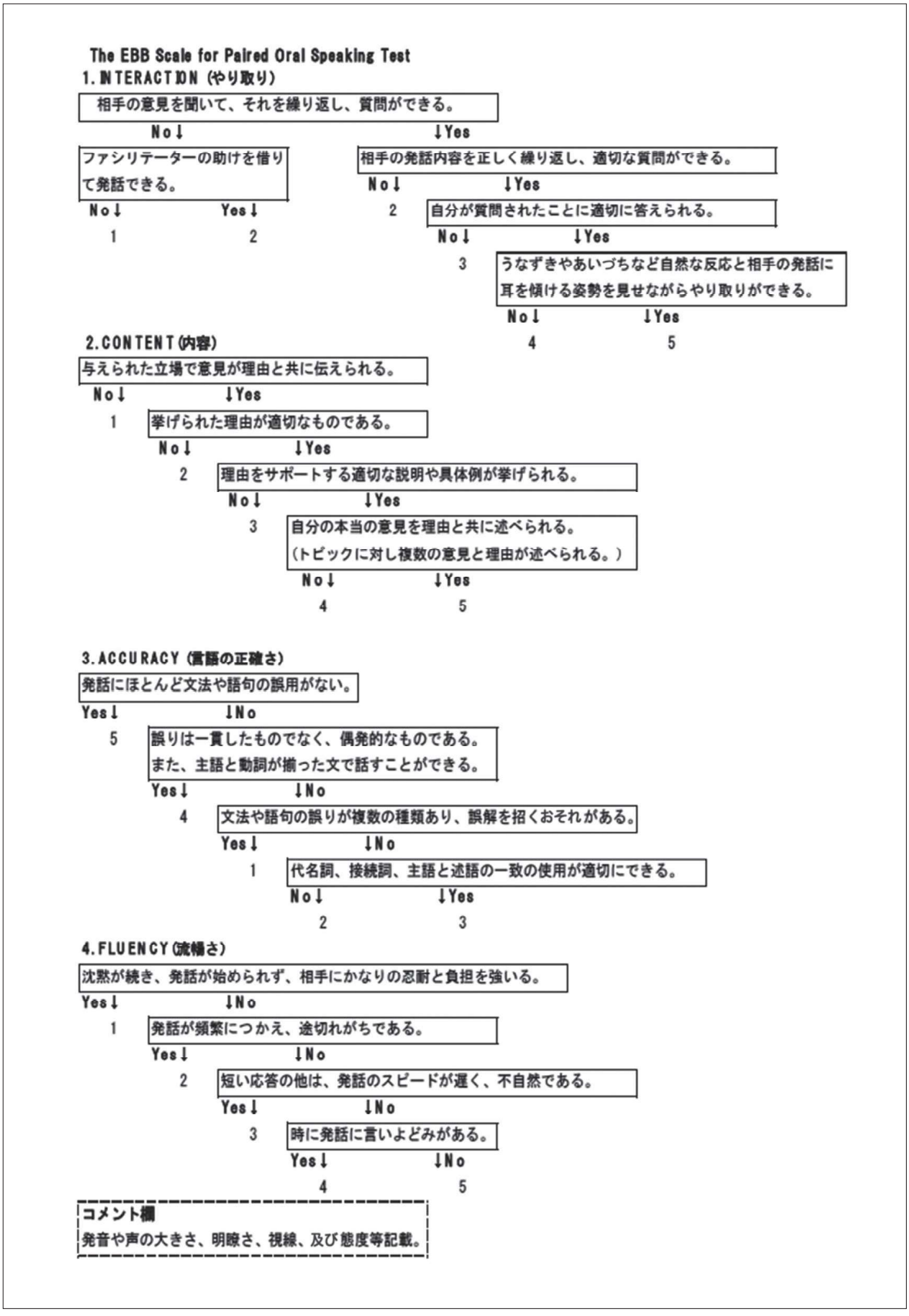
＊ Upshur, J.A., & Turner, C.E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49.*3-12. doi: 10.1093/elt/49.1.3

＊ Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23,* 411-440. doi: 10.1191/02655322061t336oa

＊ 山森光陽 (2004).「英会話テストの信頼性の検討ーー般化可能性理論ー」三浦省五（監修）前田啓明・山森光陽（編著）磯田貴道・廣森友人（著）『英語教師のための教育データ分析入門 授業が変わるテスト・評価・研究』(pp. 82-89). 東京：大修館書店

＊ Young, R.F. (2011) Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426-443). London: Routledge.

資料1： 本研究で用いたPaired Oral EBB ルーブリック（数字は得点を示す）



資料2: Paired oral 診断的フィードバックシート

やりとり Interaction /5	◎インタビューに懸命に取り組む姿勢が評価できます。	○インタビューの手順を予め確認し、練習は独り言でもよいのでロに出してみよう。
	◎先生に助けをもらいながら会話を続けることができています。	○自然なやりとりが自発的にできるように、会話の流れや手順を確認しましょう。
	◎インタビューの手順や自分の話すべきタイミングが理解できています。	○相手の言ったことをよく聞き、理由を問うための適切な質問ができるようにしましょう。
	◎相手の発言を聞いて理解し、理由を問う適切な質問ができています。	○自分の話すことだけに集中せず、相手の質問もよく聞いて、適切に応答しましょう。
	◎相手の質問を聞き、適切に応答できています。	○相手の発言している時は、頷きなど、耳を傾けているという態度を伝えてみよう。
	◎相手の話にも自然に応えられ、しっかりと自分の意見も伝えられています。	○更に自然なやりとりのために、表情やタイミング、姿勢などにも配慮してみよう。
内容 Content /5	◎課題に対して真面目に取り組み、努力する姿勢が伝わります。	○キーワードや話す内容を準備すると、リラックスし自信を持って話すことができます。
	◎課題を理解し、関連した適切なポイントを1つ挙げることができています。	○自分の挙げているポイントが一貫して適切であるかを考えながら話してみましょう。
	◎与えられた立場で課題に即したポイントを複数挙げられています。	○自分の挙げたポイントとその理由が合致して矛盾がないかを考えながら話してみよう。
	◎与えられた立場で適切なポイントと理由が共に述べられています。	○実際に自分が考えている意見をとっさに述べられるように、英語をロに出して練習しよう。
	◎与えられた立場で適切なポイントを挙げ、自分の意見と理由を述べています。	○更に自分の独自の意見も伝えられるよう、相手が納得するような理由を考えてみよう。
正確さ Accuracy /5	◎言いたいことを自分の言葉で伝える気持ちが伝わります。	○正確に話すために、まずは簡単な表現を書いて読み上げる練習からしてみましょう。
	◎不完全な文での表現もありますが、言いたいことを概ね伝えられます。	○主語や動詞、つなぎ言葉を間違えると正確に伝わらないこともあるので気を付けよう。
	◎適切な主語やつなぎ言葉を使って話すことができています。	○間違って覚えてしまった言葉や文法表現がないか、もう一度確認してみましょう。
	◎うっかりした言い間違いはあっても言いたいことを正確に伝えられます。	○簡単な表現で構わないので、慣用表現など正確に表現できるよう練習してみましょう。
	◎適切な語彙、文法を用いて正確に相手に自分の言葉を伝えられています。	○短くて簡潔な表現で良いので、少しずつ話せる語彙を増やしてみましょう。
流暢さ Fluency /5	◎真摯に伝えようと努力する姿勢が相手に伝わります。	○なかなか話し始めず沈黙すると相手が困ってしまうので、単語レベルでも発してみよう。
	◎適切なタイミングで話し始めることができます。	○途中で言葉が途切れ途切れになったり、沈黙してしまうので気を付けてみましょう。
	◎途中で止まってしまわず、自分なりのペースで話ができます。	○慌てる必要はないですが、聴き手に負担をかけないスピードでの発話を心がけてみよう。
	◎時に詰まってしまいますが、聴き手に負担をかけないペースで話せます。	○速く話す必要はありませんが、相手が安心して聞けるスムーズな発話を目指してみよう。
	◎よどみなく聴き易い発話です。	○ロに出す練習を更に重ねていきましょう。
先生からのコメント		

資料3: ペアオーラルインタビュー振り返りシート2種（紙面の都合上、記述欄は縮小して掲載）

学籍番号　　学科　　氏名

今日の日付：　　年　　月　　日

あなたは：Student A（オンラインショッピング）　・Student B (インストアショッピング)

1．インタビューを振り返ってうまくいった点

2．うまくいかなかった点

3．感想

Paired Oral（ペアで意見を交換する活動）をやってみて、その効果があった、あるいは良かったと少しでも感じる事柄を思い出して3つ書き出してください。そしてまた、その度合いを3点満点中どの程度か右の☆を塗りつぶして示してください。

	効果小	効果中	効果大
例）英語を勉強する気持ちが高まった。	☆	★	☆
①	☆	☆	☆
②	☆	☆	☆
③	☆	☆	☆

Paired Oral のフィードバックシート（先生からの評価シート）は役立ちましたか。また、その感想や意見を書いてください。

Paired Oral testは学生同士で会話をする形式ですが、英検などの面接官の先生と1対1のスピーキングテストと比べて以下の質問に答えてください。

1. あなたは今までに、英検などの外部試験あるいは学校の授業、塾や英会話学校などで、1対1のスピーキングテストを受けたことがありますか。該当するものに○をつけて下さい。
（　ある　・　ない　・　覚えていない　）
2. Paired Oral testと通常の先生と1対1の形式のテストではどちらの方が難しいと感じましたか。○をつけてその理由を書いてください。
（　Paired Oral test　・　通常のスピーキングテスト　・　どちらとも言えない　）

資料4: Paired Oral タスク2 指示文

STUDENT A

英語ペアインタビューの前に次の指示文を1分間黙読します。手順をよく確認しましょう。
指示は日本語で書かれていますが、会話は全て英語で行います。

1．先生に名前を聞かれるなどの**簡単なやりとり**をします。よく聞いて答えましょう。
2．**オンラインショッピング（ネットショッピング）の良い点をあげます。**
＜相手が質問をします＞
3．**相手の質問に答えます。**
＜答えた後、相手の反応をきちんと確認します＞
＜相手がインストアショッピング（お店での買い物）の良い点をあげます＞
4．相手の言った「インストアショッピングの良い点」を**リピートしてからそれをあげた理由を質問します。**
5．相手の答えをしっかり聞いて反応をします。
6．最後に先生の質問を聞いて、**自分の本当の意見とその理由（これまでの会話の中で出ていないもの）を述べます。**

※これでペアオーラルインタビューは終わりです。お疲れ様です。インタビュー後の感想を書いて提出してください。

STUDENT B

英語ペアインタビューの前に次の指示文を1分間黙読します。手順をよく確認しましょう。
指示は日本語で書かれていますが、会話は全て英語で行います。

1．先生に名前を聞かれるなどの**簡単なやりとり**をします。よく聞いて答えましょう。
＜相手がオンラインショッピングの良い点をあげます＞
2．相手の言った「オンラインショッピングの良い点」を**リピートしてからそれをあげた理由を質問します。**
3．相手の答えをしっかり聞いて反応をします。

4．**インストアショッピングの良い点をあげます。**
＜相手が質問をします＞
5．**相手の質問に答えます。**

6．最後に先生の質問を聞いて、**自分の本当の意見とその理由（これまでの会話の中でででていないもの）を述べます。**

※これでペアオーラルインタビューは終わりです。お疲れ様です。**インタビュー後の感想を書いて提出してください。**