〈研究論文〉

意見交換型タスクでの Paired oral テストの信頼性の検証

松村　香奈
（外国語学部英米語学科）

Dependability of Opinion Exchange Tasks Response Ratings in a Classroom-based Paired Oral Test

Kana MATSUMURA
（Department of English Language Studies, Faculty of Foreign Language Studies）

Paired oral テスト とは、面接官と受験者が対話を行う従来の面接型とは異なり、英語学習者同士が対話を行う形式のスピーキングテストである。本研究では、与えられたトピックに対し、異なる意見を述べるタイプのタスク向けの EBB 尺度（Empirically derived, Binary-choice, Boundary-definition scale; 以下 EBB 尺度とする）を作成し、多変量一般化可能性理論を用いて、テストの測定値に含まれる測定誤差の成分や大きさを分析する（Brennan, 2001a）ことで、十分な信頼性を確保するための評価者およびタスクの数をシミュレーションで検討した。その結果、決定研究（D-study）から、教室で行われる、半構造化された評価基準が明確な意見交換型タスクという条件下では、1 人あるいは 2 人の評価者でも 3 つのタスクがあれば、信頼性が確保されたテストが実行できる可能性が示された。

キーワード：Paired oral テスト、多変量一般化可能性理論、EBB 尺度、診断的フィードバック、意見交換型タスク

はじめに

Paired oral テストは、英語母語話者や英語教員などの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、受験者自身と同程度の英語力の相手と会話をすることで、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）。母国語が英語ではない者同士のペアあるいはグループでの英語スピーキング能力に対して、ヨーロッパ言語共通参照枠（the Common European Framework of Reference：CEFR）（Council of Europe, 2001）では、Interaction の評価基準が示される。共通語としての英語（English as a Lingua Franca；以下 ELF とする）視点でのスピーキング能力を測る英語検定試験では、ケンブリッジ英語検定のスピーキングテストの一部で実施されている。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 テストが活用されている。一方、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（Bonk & Ockey, 2003）妥当性、公平性に問題があると指摘する研究もある（Iwashita, 1996）。このため日本では、ELF 視点でのコミュニケーション能力の重要性の認識はあっても、妥当性、公平性の問題点を抱える中

で、利害関係が大きいテスト（high-stakes test）での実施には至っていない。しかし、テスト得点使用（test score use）に関しては、Koizumi, In'nami and Fukazawa（2016a）が指摘するように、利害関係が比較的小さいテスト（low-stakes test）では、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、先に述べたような困難はさほど問題にならないのではないかと考えられる。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は、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。

これまで日本の大学の教室内で行われる英語学習者を対象とした Paired oral テストのタスクは、Koizumi, In'nami, and Fukazawa（2016 a, b）に見られるように､仮想の場面でのロールプレイングや、Negishi（2015）での家族や学校といった身近なトピックでの自由会話であることが多く、日常会話でのやり取りを伴う英語力を測る研究がなされてきた。本研究では、与えられたトピックに対して自らの意見を理由と共に述べると同時に、異なる立場の意見を聴きながらやり取りをするタスクで Paired oral テストを実施した。これは、本研究が授業の一環として教室で行うテストで、対象科目が「授業カリキュラムの英作文を中心に、実用的な英語力の定着」を目的とし､「様々な情報を発信したり、意見を述べたりすることができる」ことが達成目標の 1 つである（Integrated English 3 共通シラバス, 2017 ; 2018 より引用）ことから設定した。

1．本研究の目的と背景

1－1　EBB 尺度の有用性

ライティングやスピーキングにおけるパフォーマンス評価の方法の 1 つに、評価基準となる記述子（descriptor）を階層的に提示し、評価者が Yes/No の二者択一の方法で評価を進めることにより得点を導くという EBB 尺度がある。EBB 尺度は、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。

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、診断的フィードバックの提供を目的とする教室での授業の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、評価尺度の実用性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。

1－2. 評価表の信頼性の検証

言語テストにおける妥当性理論の変遷および妥当性・有用性検討に関する近年の動向については、澤木（2011）をご参照いただきたい。1990 年代以降の言語テストでのテスト妥当性研究の枠組みについて、バックマン、パーマー（1996/2000）のテストの有用性の概念として紹介される 6 つの特質（qualities）[1] のうちの 1 つが信頼性である。信頼性（reliability）は､「 テストの結果が一貫したものであるか、つまり、同じ受験者がいつ、どこで受けて誰が採点しても同じ結果が得られるものであるか」と定義される（澤木, 2011, p.55）。信頼性は､「テストの得点解釈の前提となるもの」であり、持続可能なテストを実施する上で基本となる事項である。これは大規模テストだけでなく、教室で授業の一環として行われる比較的利害関係が少ないテストであっても、評価およびそれに基づく学習者への適切なフィードバックを提供するという観点からも一定程度の信頼性が確保されるべきである。本研究に用いた評価表である EBB 尺度の信頼性について、多変量一般化可能性理論を用いて検証した。

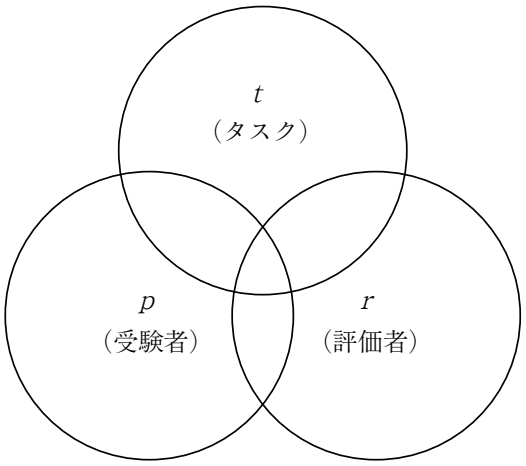
1－3. 多変量一般化可能性理論による分析

山森（2004）にあるように、一般化可能性理論とは、分散分析（ANOVA）の原理に基づき、学力やアンケート結果などに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。パフォーマンステストにおいては、測定値に含まれる誤差が、受験者の能力差に加え、評価者やタスクの違いなど複数、複合的な事柄に起因する（**表**

1)。その測定誤差が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、いくつのタスクが必要かをシミュレーションで検討できる。本手法のメリットは、人的、時間的制約のある教室でのテスト実施に有用な情報を提案できる点にある。いわゆる信頼性係数[2]にあたる数値には、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[3]である（図 1）。本研究では、単変量ではなく多変量一般化可能性理論を用いることで、複数の観点での分析的評価を分析、観点間の比較や相関を検証した。分析ソフトは、mGENOVA（Brennan,

表 1　各変動要因における推定成分の解釈

変動要因	推定分散成分の解釈
Person（*p*）	受験者の能力を識別できる割合
Raters（*r*）	評価者間での評定の厳しさのばらつき
Tasks（*t*）	タスク間でのタスクの難易度のばらつき
pr	評価者によって受験者の順位が入れ替わる程度
pt	タスクによって受験者の順位が入れ替わる程度
rt	評価者とタスクの交互作用
prt + e	全要因の交互作用＋その他の誤差



Interaction					Content				Accuracy				Fluency			
*r*1（評価者 1）			*r*2（評価者 2）		*r*1（評価者 1）		*r*2（評価者 2）		*r*1（評価者 1）		*r*2（評価者 2）		*r*1（評価者 1）		*r*2（評価者 2）	
p	*t*1	*t*2	*t*1	*t*2	*t*1	*t*2	*t*1	*t*2	*t*1	*t*2	*t*1	*t*2	*t*1	*t*2	*t*1	*t*2
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

図 1　本研究における実験計画の概念図（*p*・×*r*・×*t*・デザイン）

2001b）を用いた。

2．研究課題

本研究は、英語を専門としない本学の大学生を対象とした教室での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、テストの実行可能性、評価尺度の信頼性、実用性を調査することを目的とし、以下の 3 つの研究課題を設定した。

(1) Paired oral テストに用いた EBB 尺度は、評価表として適切に機能しているか。
(2) EBB 尺度の 4 つの分析的評価観点の結果から何が分かるか。また、観点間の関係性はどのようものか。
(3) 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。

3．研究方法

3－1. 参加者

2018 年 4 月から 7 月にかけて、英語 4 技能の育成を目的にした必修科目（Integrated English 3）の 3 クラスの受講者の内、スピーキングテストを受験し、承諾書を提出した 69 名を分析対象とした。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、2 級：5 名、準 2 級（あるいは TOEIC Bridge130 前後）:20 名、3 級:28 名、不明：15 名である。全員母国語は日本語で、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：1 名、大学入学前に英国に通算 3 年：1 名）。

なお、研究実施に当たっては、法人で定める個人情報の保護の規則に則り、倫理的配慮に十分留意した。調査協力者に対して、授業最初でのオリエンテーションの際に、調査目的、データの利用に関して口頭及び文書での説明を行った。研究結果は、個人が特定されないかたちであることを説明した上で承諾書を提出してもらった。分析に当たっては、承諾書の提出がなかった場合には分析対象から除外された。データは全て暗号化され、個人の特定化がされない形で行われた。

表２　参加者概要

科目名	種別	学生数（男 : 女）	学年	専攻
授業 A	必修	29（13：16）	2 年	経営学
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

3－2. 使用タスク

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、自由会話や文脈が設定されたロールプレイ型のタスクに比較して、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指示文により半構造化できるため、レベルの相違による発言の機会の極端な偏りや、沈黙を避けることができる。採点項目が学生と評価者に明確であることで、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せる。同時に、明確な評価項目は評価の信頼性に貢献でき、フィードバックの焦点が明確になる。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、テキストに関連する 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筆者が選んだ。

3－3. 評価観点及び尺度

EBB 尺度は、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、作成の経緯の特徴から、達成できたこととできないことが明確になり、診断的評価に適していると述べる。本研究では、先行研究（Koizumi *et al.* ,2016a ; 2016b）を参考に、やり取り（Interaction）、内容（Content）、正確さ（Accuracy）および流暢さ（Fluency）の 4

つの観点で EBB 尺度を作成した（**資料 1**）。

3－4. 研究実施文脈

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、テキストを用いたライティングスキル育成を目的としたコミュニケーション活動を、個別活動では、Paired oral テストの実施と個人のライティング活動とした。全体活動は、Paired oral テストの形式に慣れるため、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書き､述べるなど英語 4 技能を統合した活動とした。Paired oral テストは準備が整った第 6 週目から始め、授業時間後半 20 分～ 30 分を使い、複数のペアに著者が Paired oral テストを行い、他の学生はライティングの課題を進めた。Paired oral は後日評価のため録音された。評価は筆者および研究協力者（第二言語習得研究を専攻する博士課程の学生）の計 2 名で行った。

3－5. 評価後のフィードバック

Paired oral テストの結果は、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 尺度をもとに作成されたフィードバックシートは、各観点での得点および総合点記入欄に加え､「達成できたこと」と「取り組むべき課題」が各観点・得点別に一覧できる。評価者が、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ことでフィードバックシートが完成し、必要に応じて個別コメントを記入する。

4．結果と考察

4－1. 研究課題 1：EBB 尺度は、評価表として適切に機能しているか

mGENOVA による分析結果は**表 3** の通りである。結果からわかることは次の 3 点である。

まず、分散成分全体に占める受験者変動要因（p）の推定分散成分の割合が 60% から 70% 近くと非常に高い割合を占めていることである。これは、評価得点のばらつきが受験者に起因する割合､すなわち、受験者の能力を識別できる割合を示すもので、この評価尺度は受験者の能力別に順位を付けられる可能性を示唆しており、これは評価尺度の信頼性で最も重要な要素となる。

次に、評価者（r）およびタスク（t）の推定分散成分が限りなく 0 に近い点である。これは、評価者やタスクによる一貫した影響がほとんど見られないことを示しており、評価者間で評価尺度の解釈あるいは採点の厳しさに大きな差異がないこと､また、全体としてタスクにより評価つまり難易度が大きく変わらないことを示し、テスト結果の一貫性という点で評価尺度の信頼性の面で貢献する結果である。

さらに､受験者（p）､評価者（r）およびタスク（t）間での交互作用については、受験者と評価者の交互作用（$p \times r$）推定分散成分が 0% ～ 6.45%、受験者とタスク（$p \times t$）が 14.53% ～ 26.81% と最も高く、評価者とタスク（$r \times t$）がほぼ 0% という結果に

表 3　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%）
$p \times r$（交互作用）	0.00（0.00%）	0.02（2.15%）	0.03（2.61%）	0.08（6.45%）
$p \times t$	0.37（26.81%）	0.22（23.66%）	0.19（16.52%）	0.18（14.52%）
$r \times t$	0.00（0.00%）	0.01（1.08%）	0.00（0.00%）	0.00（0.00%）
$p \times r \times t, e$	0.12（8.70%）	0.13（13.98%）	0.16（13.91%）	0.12（9.68%）

注）（　）内は、各観点の分散成分全体に占める各変動要因の分散成分の割合（percentage of variance explained）。
変動要因解釈については表 1 を参照 .

なった。これは、タスクによって受験者の順位が一定程度入れ替わる可能性がある程度あることを示している。すなわち、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推測される。最後に、3 つの交互作用および残差は、様々な要因が関わるが、割合が 10% 程度で大きな問題はないと判断できる。

以上の結果から、本研究で使用した EBB 尺度は、テストの結果が一貫したものであり、同じ受験者がいつ、どこで受けても、また評価者が変わっても同じ結果が得られる可能性が高く、従って信頼性の観点から、評価表として一定程度適切に機能していることが示唆されたと考える。

4－2. 研究課題 2：EBB 尺度の 4 つの分析的評価観点の結果から何が分かるか。また、観点間の関係性はどのようなものか。

表 3 の観点ごとの推定分散成分から分かる各観点の特徴は次のようにまとめられる。

Interaction（やり取り）は、他の観点に比べ、受験者とタスクの交互作用（$p \times t$）0.37 と大きく、観点内で占める割合も 26.81% と比較的大きい。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、つまり、受験者の中に特定のタスクでやり取りに関して取り組みやすさに差があることを示している。この要因の可能性の 1 つとして、タスクによっては相手の予想外の発言により、やり取りに難しさが生じている可能性が考えられる。これは、Content（内容）の観点でも同じく受験者とタスクの交互作用が大きめであることから、タスクがやり取りや発言内容に関して、特定の受験者に影響を与えていることが推測できる。

次の Content（内容）は、4 つの観点の中では一番測定誤差が小さいことがわかる。また、受験者によって差がつきにくいことも受験者（p）の値が分散成分の割合が他の観点より比較的小さい（59.14%）ことから推察できる。これは、テストのトピックを予め提示していることで、前もって話す内容を準備することが可能であることから、順当な結果であると考える。

Accuracy（正確さ）の特徴として、他の観点より評価者（r）の一貫した影響が 0.04（3.48%）と若干大きめに存在する点である。これは、評価者の正確さについては、評価者間で若干厳しさに差がある可能性が考えられる。これは、授業の指導教員が評価者の 1 人であり、受験者に対する正確さの達成目標に対する認識に評価者間で差があったことに起因する可能性もあるが、値が小さいことから信頼性に関しては問題ないと考える。

Fluency（流暢さ）は、Interaction と共に、受験者要因（p）が比較的大きく、受験者の能力測定に貢献度が高い、つまり受験者の能力差を測る観点として信頼性が高いことを示していると推測できる。

最後に、**表 4** の 4 観点間の相関行列から、各相関は 0.70 ～ 0.90 と各項目の独立性を保ちながら同時にいずれの組み合わせも 0.7 以上の高い相関性を示し、4 つの観点が共通した概念を測定しているということを示唆している。このことは、テストの構成概念妥当性[4] の観点から非常に重要な点である。また、Accuracy（正確さ）は他のどの観点とも最も相関性が高く、当然のことながら、英語で正しく表現できる力が他の観点と大きく関わることを示している。

表 4　分析的評価 4 観点間の相関行列

	Interaction	Content	Accuracy	Fluency
Interaction	1.00			
Content	.70	1.00		
Accuracy	.86	.89	1.00	
Fluency	.78	.79	.90	1.00

5–3. 研究課題 3：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。

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決定研究（D-study: decision study）[5] での分析を行った。

表 5 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。教室実施の実効性と利害関係が小さ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。表では、斜字及び下線で信頼度指数が 0.8 以上になる組み合わせを示した。

表５　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）（p x r x t 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．

決定研究の結果から、評価者が全てのタスクを評価するデザインにおいて、３タスクを評価者２で評価する場合、全ての観点で信頼度指数が 0.8 以上となり、理想的であるが、評価者が１人でもタスクが３つあれば十分に信頼性の高いテストデザインになることが示唆される。また、２タスク- ２評価者よりもむしろ信頼性が高くなる可能性が示されるが、これは、**表３**の受験者とタスクの交互作用での割合の結果から分かるように、タスクが多い方が、評価者にとって取り組み易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。

教室での評価では、ティームティーチングでない限り、現実的には指導教員が単独で評価する場合が多いと考えられるが、１人の評価でも３タスク行うことで、信頼性が確保された評価が期待できる。

５．結論・課題

本研究で用いた Paired oral テストの EBB 尺度は、多変量一般化可能性の分析結果から、評価尺度として適切に機能していると判断できた。やり取り、内容、正確さ、および流暢さの４つの観点の相関性は高く、共通の構成概念を測っていることが推測される。4 観点の中では、「正確さ」がその他の観点との相関が最も高く、正しい英語表現を身に付けることが Paired oral テストにおいても重要であると考えられる。また、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、また、評価者が全てのタスクを評価するというデザインにおいて、３タスクを２評価者で評価するのが効率的で信頼性が確保されたシナリオであるが、１人の評価者でも３つのタスクがあれば、信頼性が確保されることから、教室でのテストとしての実行可能性が示唆された。

ただし、今回の研究では、評価者２人が評価トレーニングを重ねていたことは、評価者間の測定誤差が小さくなった要因であると考えられる。また、タスク選定についても、事前に学生にアンケートで、取り組み易いトピック２つを選定したことも、タスクの差異による取り組みの違いを避ける効果があったと思われる。

結論として、Paired oral テストはこれまで、ある程度英語力の高い学習者を対象として実施されてきたスピーキングテストであるが、英語を専門としない学習者でも、レベルにあったタスクの設定と、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る EBB 評価尺度を採用することで、実用的であると同時に信頼性の確保されたテストを実施できると考える。

今後の課題としては、信頼性の検証に加え、妥当性検証のために、他のテスト得点との関係性を測る外挿（Extrapolation）の視点での検証や、受験者に対する質的分析による検証の実施が必要となると考える。

謝　辞

本研究の一部は、公益財団法人日本英語検定協会英語教育センターによる第 31 回英検研究助成をけたものである。

注

1）テストの有用性は、(1) 信頼性（reliability），(2) 構成概念妥当性（construct validity），(3) 真正性（authenticity），(4) 相互性（interactiveness），

（5）影響（impact），（6）実用性（practicality）の 6 つの特性から成る（澤木，2011，p.55）。

2）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）。

3）受験者（p: person）、評価者（r: rater）、タスク（t: ask）の 3 相完全クロス計画（$p \cdot \times r \cdot \times t \cdot$）である。

4）構成概念妥当性とは、「テストの得点を、どの程度そのテストで測りたい構成概念の指標として解釈することが可能か」（澤木, 2011）というもの。

5）決定研究とは、タスクの数や評価者数を変えながら多変量一般化可能性係数の変化を予測することによって、具体的な評価実施についての予想情報を引き出し、一定の条件下での評価実施計画シナリオを考える方法である（池田，1994）。

参考文献

池田央（1994）『現代テスト理論』，朝倉書店 .

澤木泰代（2011）「大規模言語テストの妥当性・有用性検討に関する近年の動向」，『言語教育評価研究』第 2 号, pp. 54−63.

バックマン . L. F., パーマー . A. S.（2000）. 大友賢二・ランドルフ・スラッシャー（監修）『実践言語テスト作成法』大修館書店.（Bachman, L.F., & Palmer, A.（1996）. *Language testing in practice*. Cambridge, UK: Cambridge University Press.）

山森光陽（2004）.「英会話テストの信頼性の検討−一般化可能性理論−」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 .

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.

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）.

Fulcher, G.（1996）. Testing tasks: issues in task design and the group oral. *Language Testing 13*（1），23-51.

Fulcher, G., Davidson, F., & Kemp, J.（2011）. Effective rating scale development for speaking tests: Performance decision trees. *Language testing, 28*（5）−29. doi:10.1177/02265532209359314

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.

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

Knoch, U.（2009）. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*（2），275-304.

Koizumi, R., In'nami, Y., & Fukazawa, M.（2016a）. 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

Koizumi, R., In'nami, Y., & Fukazawa, M. (2016b). 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).

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.

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

Swain, M. (2001). Examining dialogue: 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

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 cgroup oral test. *Language Testing, 23*, 411-440. doi: 10.1191/0265532206lt336oa

(受付日：2019年10月30日、受理日2019年12月27日)

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）

The EBB Scale for Paired Oral Speaking Test

1. INTERACTION（やり取り）

相手の意見を聞いて、それを繰り返し、質問ができる。

↓Yes

相手の発話内容を正しく繰り返し、適切な質問ができる。

↓Yes

自分が質問されたことに適切に答えられる。

↓Yes

うなずきやあいづちなど自然な反応と相手の発話に耳を傾ける姿勢を見せながらやり取りができる。

↓Yes　5

No↓　4

No↓　3

No↓　2

No↓　ファシリテーターの助けを借りて発話できる。

Yes↓　2

No↓　1

2. CONTENT（内容）

与えられた立場で意見が理由と共に伝えられる。

↓Yes

挙げられた理由が適切なものである。

↓Yes

理由をサポートする適切な説明や具体例が挙げられる。

↓Yes

自分の本当の意見を理由と共に述べられる。
（トピックに対し複数の意見と理由が述べられる。）

↓Yes　5

No↓　4

No↓　3

No↓　2

No↓　1

3. ACCURACY（言語の正確さ）

発話にほとんど文法や語句の誤用がない。

↓No

誤りは一貫したものでなく、偶発的ななものである。
また、主語と動詞が揃った文で話すことができる。

↓No

文法や語句の誤りが複数の種類あり、誤解を招くおそれがある。

↓No

代名詞、接続詞、主語と述語の一致の使用が適切にできる。

↓Yes　3

No↓　2

Yes↓　1

Yes↓　4

Yes↓　5

4. FLUENCY（流暢さ）

沈黙が続き、発話が始められず、相手にかなりの忍耐と負担を強いる。

↓No

発話が頻繁につかえ、途切れがちである。

↓No

短い応答の他は、発話のスピードが遅く、不自然である。

↓No

時に発話に言いよどみがある。

↓No　5

Yes↓　4

Yes↓　3

Yes↓　2

Yes↓　1

コメント欄

発音や声の大きさ、明瞭さ、視線、及び態度等記載

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口に出してみよう。
	◎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、英語を口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。	○口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		