様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ９ 月 １６ 日現在

機関番号：３２７１０
研究種目：研究活動スタート支援
研究期間：2013～2014
課題番号：２５８８４０７３

研究課題名（和文）英語口頭試験の形式および対話者の言語レベルの違いは評価に影響するか

研究課題名（英文）Effects of test types and interlocutors' proficiency on oral performance assessment

研究代表者
根岸　純子（Negishi, Junko）

鶴見大学・文学部・准教授

研究者番号：１０７０８９６０

交付決定額（研究期間全体）：（直接経費）　　1,300,000円

研究成果の概要（和文）：本研究では、英語力の異なる24名の大学生を対象に、単独・ペア・グループの３形式の口頭模擬試験および評価を実施し、様々な角度から分析を行った。分析の結果、被験者は単独形式よりもグループ形式で高得点、ペア形式で低得点となる傾向が見られたことから、英語能力測定には複数被験者形式のみではなく、単独等の他形式試験と組み合わせる必要性が示唆された。今後、複数被験者形式が日本に導入される可能性とインタラクション能力の育成という点を鑑み、対話型の言語活動を教室内で増やすことも必要であろう。

研究成果の概要（英文）：The study compared the impacts of three types of test (a single-speaker task, paired oral discussion, and group oral interaction) and interlocutors' proficiency level, on the test scores of 24 Japanese university students. The results showed that the students scored higher in the group oral interaction compared to the single-speaker task and paired oral discussion. Paired or group oral interactions may be implemented in test batteries in the near future. Therefore, learners may need more communicative, interactive activities in their classrooms, to promote their interactive competence.

研究分野：人文学

キーワード：外国語教育　第二言語習得　テスティング　スピーキング　評価　インタラクション

１．研究開始当初の背景

(1) インタラクション能力測定

近年、外国語によるコミュニケーション能力の測定を目的とし、実際に言語を話す場面を設定した口頭試験が増えてきている。これは、現代のグローバル化した世界において、第二言語あるいは外国語学習者の実践的な会話力を測定する必要性が増大しているからに他ならない。英語学習者のスピーキング能力を評価する際、単独話者を対象とすることが多いが、これは様々な要因に影響されることなく話者の能力を引き出すことができるからである。しかし、単独話者のパフォーマンス（与えられた絵を説明したり、マイクに向かって話をしたりする、一人だけの言語産出）は、現実の会話で起こり得る意味の確認や交渉などがなく、自然な会話とは言い難い。二人で対話をする面接は、より現実の会話に近いが、面接官が質問することに被面接者が答えるという非対称な力関係の下で実施されていること等が問題視されている。

それに比較すると、ペアやグループによるインタラクション（相互対話）においては、話者が与えられた状況下で動的なインタラクションを構築できること（Swain, 2001)、実践的であること、実際の教室でのスピーキング活動を活用できること、教育機関などでは、より少ない時間でより多くの被験者を対象に教師が評価に専念できること等を利点として挙げることができる。このように、ペアやグループのインタラクションに多くの利点がありながら、これまで口頭試験として利用されてこなかった理由としては、会話をする相手の性格や言語レベルに少なからず影響を受けてしまうこと、評価者間信頼性が低いこと（Van Moere, 2006）等の欠点が挙げられている。

(2) 各国の状況

単独発話や複数話者間のインタラクションには、それぞれ利点・欠点が存在するが、近年、諸外国においては、ペアやグループ形式の口頭試験が導入されるようになっている。例えば、ケンブリッジの実施する２種類のテストでは単独発話試験に加えてペアやグループによる口頭試験が取り入れられている。また、欧州協議会（Council of Europe, 2001）の作成した「欧州言語共通参照枠（Common European Framework of Reference: CEFR)」を使用したペアおよびグループによる口頭試験も実施されている。アジアにおいても、香港や中国ではグループ形式の口頭試験が実施されており、韓国では、グループによる口頭試験が大学の入学試験や奨学生選考にも利用されている。これらの実情を鑑みると、近い将来、日本でも複数話者による口頭試験が実施される可能性が高いと思われる。

(3) 複数話者による口頭試験と研究

ペアやグループによる口頭試験は、比較的新しく導入された形式であり、日本はもとより、世界においてもその研究は緒に就いたばかりであると言える。世界的にみると、ペアの口頭試験については、最近、研究が進められるようになってきているものの、グループによる口頭試験に関しての研究は、ペアに比較して非常に少ないのが現状である。

２．研究の目的

(1) 研究課題

研究よりも実践が先行している状況下、単独話者およびペアとグループのインタラクションという３形式の口頭パフォーマンス試験を比較した研究が未だ行われていない。そこで、英語力の異なる日本人英語学習者を対象にこれらの口頭模擬試験を行って評価をし、以下の研究課題について分析することが本研究の目的である。

(A) 異なる３形式の口頭試験（単独・ペア・３名グループ）は評価に影響を与えるか。

(B) 複数話者を同レベルあるいは異なるレベルで組み合わせた場合、評価に影響を与えるか。影響を与える場合、すべてのレベルで同様の影響がみられるか。

(C) 複数話者を評価する場合、評価者は同一グループの構成員に同一評価をする傾向はあるか。

(2) 期待される成果

上記の課題について分析することにより、複数話者対象の口頭試験を行い、単独話者の場合と同様の結果を得ることが可能であるか、また、複数話者の組み合わせを行う際に、ある程度予想される言語能力による組み合わせが必要であるのか、あるいはランダムに組み合わせが可能であるかについての知見が得られる等、将来この形式を取り入れる際の参考とすることができると期待される。

３．研究の方法

(1) 被験者

被験者は２大学計 24 名の大学生で、全員が３形式の口頭模擬試験に参加した。A大学の 12 名は、外国の大学とインターネットを利用した異文化交流を選択履修している学生で、毎週３時間、英語でのプレゼンテーションおよび討論を行っている。学部は様々で英語専攻者はいないものの、帰国子女や留学経験者を含む英語力の高い学生が多い。B大学の 12 名は英語専攻で１名の留学経験者を含んでいる。全体の男女数はそれぞれ 11 名および 13 名､TOEIC の点数は 300 点から 960 点であり、英語力が幅広く分布するように選定した。さらに、O'Sullivan (2002)のいう親密度による評価差を避けるため、各大学内において級友同士を組み合わせた。

(2) 倫理審査

本研究の実施に当たっては、研究者の所属する機関の倫理審査委員会による承認後、各被

験者に研究計画を説明し、同意書に署名を受けた後、実験を開始した。

(3) 言語データ収集

①単独話者による言語産出には、(財) 日本英語検定協会の実施する英語検定準１級２次試験の過去問題を、協会より許可を得た上で使用した。これは、４コマの絵の示す短い物語を一人で描写するものである。絵を見て１分間物語を考え、その後２分以内で言語産出を行った。発話時間の違いは、計算によって補正した。最初に、全員がこの単独発話に参加した。

②ペアによるインタラクションは、被験者全員が二度行った。上記①の結果・担任の授業評価・TOEIC の点数を勘案し、一度目はほぼ同レベルの言語能力の被験者同士をペアとし「家族」というトピックで、二度目は異なるレベルの被験者同士をペアにし「学校」というトピックで、計 200 秒間の対話に参加させた。

③３名グループのインタラクションも、被験者全員が二度行った。ペアと同様、３名をほぼ同レベルで組み合わせて「夢」というトピックで、次に３名のうち１名ないし２名が異なるレベルになるよう組み合わせて「英語」というトピックで、計 300 秒間、言語産出に参加させた。

以上の①、②、③は、被験者の同意を得た上でビデオ撮りをし、後述の評価および分析に供した。被験者は、ビデオ撮りの最初に口慣らしを兼ねて自己紹介をしているが、評価用ビデオを作成する際には、個人の特定できる部分を削除し、グループ番号及び着席位置のみによる表示に変えた。

(4) 評価

評価は、英語教育に 10 年以上携わっている修士以上の学位を有する、現職の日本人英語教員５名が行った。日本人評価者採用の理由は、英語教育に精通した複数英語母語話者の獲得が難しいこと、および非母語話者でも母語話者と同等の評価ができるという報告（Kim, 2009）による。

評価基準は欧州言語共通参照枠（CEFR）の日本版である、CEFR-J を使用した（バージョン 1.1；投野, 2013)。オリジナルの CEFR は 2001 年、欧州評議会が多言語・多文化環境下で、様々な言語研究をもとに作成した、スピーキングを含む言語全体に関する枠組みである。CEFR には共通参照レベルと呼ばれる、学習者の能力を示す基準があり、基礎段階の言語使用者を表すレベルＡには A1・A2、 自立した言語使用者であるレベルＢには B1・B2、熟達した言語使用者であるレベルＣには C1・C2 があり、全体として６レベルに分かれている。

一方、CEFR-J は CEFR 同様 CAN-DO（～できる）という能力記述文により、学習者が学習目標を知ることができるようになっている。現在、文科省は各中学校・高等学校において学校の実情に応じた CAN-DO リストの作成を促している (2014)。CEFR と CEFR-J の最大の相違点は、後者が日本人英語学習者に適合するよう、評価レベルを細分化していることである。これは、日本人英語学習者の８割がレベルＡに分類されることによる（Negishi, 2011)。その結果、CEFR-J では、「A1 以下」の他、レベルＡが A1.1、A1.2、A1.3、A2.1、A2.2 の５段階、レベルＢが B1.1、B1.2、B2.1、B2.2 の４段階に細分化されている。一方レベルの高いレベルＣは、CEFR 同様 C1・C2 の２レベルである。また、CEFR では、「共通参照レベル：話し言葉の質的側面」において「使用領域の幅（range)」「正確さ（accuracy)」「流暢さ（fluency）」「やりとり（interaction)」「一貫性（coherence)」という言語使用カテゴリーが５つ提示されているが、CEFR-J ではこのようなカテゴリーは存在しない。その関係で、本研究では、CEFR-J を用いた全体評価のみを分析対象とした。

本研究では、①の単独話者による言語産出の評価は CEFR-J の「話すこと：発表」、②と③の複数話者による言語産出の評価は同「話すこと：やりとり」を使用した。

評価者５名中４名は以前に CEFR を用いた評価訓練を受け、ペア・グループ対話者を評価した経験がある。今回は新たな評価者１名を含め CEFR の訓練用ビデオ(North & Hughes, 2003）を使用して訓練を行った。その後、同様の別データを用いた試験的な評価を行い、話し合いおよびすり合わせをした上で、本研究の評価作業を実施した。

(5) 分析

分析は主に量的側面から行い、一部、多相ラッシュ分析を実施した。第二言語による口頭試験においては、たとえ厳密な評価基準があったとしても、人間が評価を行うことから、主観が入ったり一貫性が損なわれたりすることもある。評価者に対する訓練は必須で、その結果、不規則なエラーを減少させたり、個人内の一貫性を向上させたりすることはできるものの、評価の厳しさ・甘さを完全になくすことは不可能である、といわれている（McNamara, 1996)。評価者の違いや課題の難しさ、被験者の能力等の複数の相 (facets)、およびそれぞれの相互作用が評価に与える様々な影響を可能な限り相殺し、評価者の出した素点を**測定値**として推定することができるのが、多相ラッシュ分析である。その結果計算されたデータにより、評価者が言語テスト形式の違いや、同等あるいは異なるレベルの話者の組み合わせによりどう評価しているのかをみることができる。多相ラッシュ分析は「FACETS」というソフトウェア（バージョン：3.7.1.4; Linacre, 2014）を使用して行った。また、素点を用いて分析した結果についても、以下に随時報告する。

４．研究成果

(1) 評価および評価者

CEFR-J を使用して評価した結果、最も下位の「A1 以下」および最も上位の「C2」と判定された被験者は皆無であったため、その２つを除き、A1.1 を 1、A1.2 を 2、…C1 を 10 というように数字に置き換えて統計処理を行った。まず、評価者について多相ラッシュ分析を実施した。その結果、評価者間に差はみられたものの、評価が一定過ぎたり、一貫性がなかったりという問題はないという結果が得られた。そのため、本分析を行うこととした。

(2) 研究課題(A)：異なる３形式の口頭試験は評価に影響を与えるか

①素点による分析

表１は３形式の口頭試験の記述統計である。全被験者はペアとグループの口頭試験を各２回ずつ受けているため、総数（試験数×被験者数×評価者数）が 240 と単独発話の倍となっている。

表１　３形式口頭試験の記述統計

形式	総数	合計	平均	標準偏差
単独	120	658	5.48	2.58
ペア	240	1266	5.28	2.54
ｸﾞﾙｰﾌﾟ	240	1341	5.59	2.48

素点の平均をみてみるとペアが最も低く、グループが最も高くなっており、単独が中間に位置している。標準偏差（SD）はいずれも 2.5 前後で、比較的大きいことを示している。分布図を作成した所、正規分布を示していなかったため、次項②では多相ラッシュ分析用ソフトウェアを使用してデータを正規化してから分析した。

②ラッシュ分析による３試験形式レポート

多相ラッシュ分析によって得られた補正値および測定値は、単独が 5.61 と－0.02(logit という単位で表している)、ペアが 5.27 と 0.12、グループが 5.78 と－0.10 となった。最も甘く評価されているのが負の測定値で表されている「グループ」で、最も厳しく評価されているのが正の測定値で表されている「ペア」であった。ただ、測定値の差は 0.22 でさほど大きくなく、エラーが 0.05～0.08 と小さいことから、差は小さいということが示された。また、適合度合いからみて、この３形式の試験はラッシュ・モデルに適合していた。しかし、カイ２乗検定の結果（$\chi^2 = 8.4$、自由度 2、$p = .01$）から、異なる３形式の口頭試験は評価に影響を与えるとの結論が導き出された。

(3) 研究課題(B)：複数被験者を同レベルあるいは異なるレベルで組み合わせた場合、評価に影響を与えるか。影響を与える場合、すべてのレベルで同様の影響がみられるか。

被験者には３(3)の②・③にあるように、上位・中位・下位の３レベルに分けて組み合わせを替えて発話活動を行わせ、評価点を分析した。

①ラッシュ分析による５種類のレポート

ここでは、３試験形式と組合せによる違いを、(ⅰ)単独(ⅱ)同レベルによるペア(ⅲ)異なるレベルによるペア(ⅳ)同レベルによるグループ(ⅴ)異なるレベルによるグループ、の５種類に分け、ラッシュ分析をした結果を示す。分析によって得られた補正値および測定値（単位：logit）は、(ⅰ)単独が 5.61 と－0.03、(ⅱ)同レベルによるペアでは 5.08 と 0.20、(ⅲ)異なるレベルによるペアでは 5.46 と 0.04、(ⅳ)同レベルによるグループは 5.81 と－0.12、(ⅴ)異なるレベルによるグループは 5.75 と－0.09 という結果であった。つまり、最も易しい試験形式（高得点を取りやすい）は(ⅳ)同レベルで組み合わせたグループ形式、次が(ⅴ)異なるレベルで組み合わせたグループ形式であった。他方、最も難しい試験形式（低得点になりやすい）は(ⅱ)同レベルで組み合わせたペア、次に低いのが(ⅲ)異なるレベルで組み合わせたペアで、(ⅰ)単独は全ての真ん中に位置していた。最も易しい試験形式と最も難しい試験形式の測定値の差は 0.32（logits）で大きいとは言えなかった。測定値の標準偏差も 0.13 で偏差は小さいと言える。標準エラーも 0.08 logits で小さい。しかしながら、前項同様、カイ２乗検定の結果（$\chi^2 = 10.7$、自由度 4、$p = .03$）から、この５種類の試験形式が同等であるとは言えないとの結論が導き出された。

これらの結果から、課題(B)の結論は、複数被験者を同レベルあるいは異なるレベルで組み合わせた場合、評価には影響を与えると言える。

②影響はどのレベルにみられるか

表２は、ペアの各被験者が得た得点と各被験者の単独発話時の得点とを比較した場合の得点変動をレベル別・組み合わせ別に表したものである。表３は同様にグループと単独発話を比較したものである。

表２　ペア発話と単独発話の評価変動

	ペアが高得点		ペアが低得点		同点
	同ﾚﾍﾞﾙ	異ﾚﾍﾞﾙ	同ﾚﾍﾞﾙ	異ﾚﾍﾞﾙ	
上　位	12.5%	0.0%	**25.0%**	**56.2%**	6.3%
中　位	**25.0%**	**49.9%**	6.3%	12.5%	6.3%
下　位	6.3%	18.8%	37.4%	31.2%	6.3%

表３　グループ発話と単独発話の評価変動

	ｸﾞﾙｰﾌﾟが高得点		ｸﾞﾙｰﾌﾟが低得点		同点
	同ﾚﾍﾞﾙ	異ﾚﾍﾞﾙ	同ﾚﾍﾞﾙ	異ﾚﾍﾞﾙ	
上　位	6.3%	12.5%	**12.5%**	**62.4%**	6.3%
中　位	**18.8%**	**68.7%**	0.0%	12.5%	0.0%
下　位	12.5%	25.0%	25.0%	37.5%	0.0%

表２および表３から言えることは、ペアあるいはグループ発話において、上位レベルの被験者が異なるレベル（中位・下位）の被験者と組んだ場合、それぞれ 56.2%、62.4%の被験者が単独発話時と比較して低得点となっていた。組み合わせ相手のレベルを考慮しなければ、それぞれ４分の３程度かそれ以上（ペア：25.0%＋56.2%、グループ：12.5%＋62.4%）がより低得点となっている。一方、中位レベルの被験者が異なるレベルの被験者と組んだ場合（実際には上位レベルの被験者と組んだ場合の方が多い）、それぞれ 49.9%、68.7%の被験者がより高得点を獲得していた。組み合わせ相手のレベルを考慮しなければ、それぞれ４分の３程度かそれ以上（ペア：25.0%＋49.5%、グループ：18.8%＋68.7%）がより高得点を得ている。一方、下位レベルの被験者には上位レベルや中位レベル被験者ほどの顕著な傾向はみられないものの、より低得点の傾向がみられた。これらの結果から、プラスの影響は中位レベルの被験者に対して、マイナスの影響は上位レベル（一部下位レベル）の被験者に対してみられると言える。

(4) 研究課題(C)：複数被験者を評価する場合、評価者は同一グループの構成員に同一評価をする傾向はあるか。

Brooks (2009) のペアでの研究、Negishi (2011)のグループでの研究では、評価者はペアやグループ内の２名あるいは３名に同じ評価をする傾向があると報告されている。そこで、ここでは、本研究のペアやグループ内被験者に対し、評価者が同一評価をする可能性が高いかどうかを分析した。

まず、ペアやグループ被験者の単独発話時の得点を用い、それを実際のペアやグループ発話時の得点と比較した。それによると、単独発話時には同得点だったペアは 7.1%であったが、実際にペアになると 15.4%に同一評価が与えられ、グループでは 13.3%だったものが 22.1%の同一評価に増加していた（但し、グループでは３名中２名のみが同得点の場合を含む）。つまり、評価者は単独発話の場合と比較し、ペアでは 2.18 倍、グループでは 1.66 倍、同得点を与える傾向があることが分かった。複数被験者の場合は同一評価をすべきである（May, 2009）という考え方もあるが、評価者は必要に応じて同一評価を与える傾向があるため、すべて同一評価とする必要はないと思われる。

(6) 結論

①結論

課題に対する分析の結果、複数話者による口頭試験は、単独話者のそれと同じ結果を得ることはできないということが明らかになった。しかしながら、複数話者による口頭パフォーマンスは、話者間の協同的インタラクションが意味の交渉を生み、その結果言語習得を促進すると言われている。このことから、複数話者による口頭試験には欠点があることは認識しつつも、今後実施される場が増えると考えられる。そのことを考えると、複数話者による発話の研究は更に進められるべきであろう。また、教師はすべてのレベルの学習者がインタラクション能力を十分に発揮できるよう、教室においてペアやグループによる言語活動を活発に進めていく必要があると考える。

②本研究の限界と今後の研究

本研究の限界は、まず、被験者数が十分ではないことである。評価者の負担と評価の質を考えての 24 名であったが、さらに大規模な産出データの取得が可能であれば一般化できる可能性がある。次に、研究が主に量的なものであったことである。今後、発話の質的研究が必要であると考えられる。また、評価基準については、複数話者に限った、実践的評価基準の作成も、今後視野に入れることとしたい。

<引用文献>

①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』大修館書店.
②文部科学省（2014).「今後の英語教育の改善・充実方策について報告～グローバル化に対応した英語教育改革の五つの提言～」http://www.mext.go.jp/b_menu/shingi/chousa/shotou/102/houkoku/attach/1352464.htm 2014 年 11 月 14 日アクセス
③Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing 26*(3), 341 -366.
④ Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.
⑤Kim, Y. (2009). An investigation into native and non-native teachers ' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.
⑥ Linacre, J. M. (June, 1997). *MESA research note #2.* Midwest objective measurement seminar, Chicago. Retrieved February 16, 2009, from http://www.Rasch.org/rn2.htm.
⑦ Linacre, J. M. (2014). *Facets Rasch measurement computer program, version 3.7.1.4.* Chicago: Winsteps.com.
⑧ May, L. (2009). Co-constructed interaction in a paired speaking test: The rater ' s perspective. *Language Testing, 26*(3), 397-421.
⑨McNamara, T. F. (1996). *Measuring second language performance.* New York: Longman.
⑩Negishi, J. (2011). *Characteristics of*

group oral interactions performed by Japanese learners of English. Unpublished doctoral dissertation. Waseda University.
⑪ North, B. & Hughes, G. (2003). *CEF illustrative performance samples: For relating language examinations to the Common European framework of reference for languages: Learning, teaching, assessment (CEF)*. Eurocentres and Migros Club Schools.
⑫ O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.
⑬Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302.
⑭Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411-440.

５．主な発表論文等

〔雑誌論文〕（計２件）

① Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *Annual Review of English Language Education in Japan (ARELE), 26*, 333-348. 査読有り
URL: http://www.jasele.jp/arele/

② Negishi, J. (In Press). Assessment behavior and perceptions of raters in paired and group oral interaction. *Journal of Pan-Pacific Association of Applied Linguistics, 19*(1). 査読有り
URL: http://paal.kr/journals/browse.html

〔学会発表〕（計２件）

①Negishi, J. (August 17th ～19th, 2014). Variability in assessment of speakers tested on three types of oral activity. *The 19th Conference of Pan-Pacific Association of Applied Linguistics*. Waseda University, Tokyo.

②根岸純子（2014 年 8 月 9 日～10 日）「英語口頭試験における試験形式および対話者の言語レベルと評価」第 40 回全国英語教育学会徳島研究大会. 徳島大学（徳島県徳島市）.

６．研究組織

(1)研究代表者

根岸　純子（NEGISHI, Junko）
鶴見大学 ・ 文学部 ・ 准教授
研究者番号：10708960