

Exploring Pair or Group Oral Discussion Oral Tests for L2 Learners

L2 学習者のための ペア/グループ・ディスカッションによる オーラル・テストの開発

担当：D.K.

1. 授業で学んだことから

Folse (2006) 第 6 章では、スピーキング能力の **assessment** について、現在までの成果の整理と今後の課題をまとめている。まず、Folse はスピーキング能力の測定の必要の理由として、**assessment** と **instruction** は循環的であり、**assessment** を行い学習者のスピーキング能力を正しく測定することにより、初めて学習者の熟達度や到達度、その学習プロセスを把握、指導に生かせると述べている。また、**assessment** について考えることは、スピーキングの構成概念やどの構成概念がどの程度必要かという程度などスピーキングの性質を明らかにすることにつながると考えられる。

Folse はスピーキング能力の **assessment** の実施にあたり、まず始めに **Need analysis** が必要であるとしている。学習者のニーズを知ることは、教師に何をどのように教えるべきかの指針を示す。ここで、日本の英語を学ぶ中高生にどのようなスピーキングのニーズがあるかを検討する。まずは、学業の面で異文化相互理解のために英語を用いて情報発信を行うためのスピーキング能力が必要である。海外の学生と意見交換をするためには、英語でのディスカッションする力が必要になる。海外へ留学する、もしくは海外からの留学生を受け入れる場合には、これらの能力に加え、日常生活場面での英語を運用することになる。また、職場に出てからも海外で研修を受けるにあたり英語で考えを述べることや、外国企業との会議・交渉・連携の場面で、スピーキング能力は不可欠なものである。つまり、中高で英語を学ぶ学習者が必要とする、あるいは将来的に必要とする能力として、英語での日常での会話能力、プレゼンテーション能力、ディスカッション能力が求められる。

では、どのようなこれらのスピーキング能力には、どのような構成概念があるだろうか。Folse はスピーキング能力評価の主な観点として、**accuracy** と **fluency** を挙げている。相手に理解可能な情報伝達をするために、音声面 (e.g., 発音や強勢) での **accuracy** と、文法面での **accuracy** は妥当であり、また **fluency** も円滑に情報発信あるいは相手とやりとりをするために妥当である。プレゼンテーションのためには、発話の内容も重要な位置を占めるために、発話の内容

的・言語的一貫性も要素のひとつであるだろう。また、ディスカッションのためには、情報や自分の考えを発信し合うだけでなく、相手との会話・議論を協同的に発展させることが必要不可欠である。この構成概念については、Bachman (1990) の言語のモデルの中に「方略的能力」が含まれることによっても支持される。方略的能力には、相手とのコミュニケーションなやりとりの目標達成のために、何が必要か判断し、相手と自分の共通する知識等を把握し、コミュニケーションの成功度を随時評価する、などの相手との「やりとり」を成功させるための能力が含まれる。

英語学習者のニーズと必要な構成概念は上記のように考えられるが、現状のスピーキング・テストは、これらのニーズを反映しているだろうか。日本の EFL 学習者が受験する機会の多い TOEFL や IELTS、英検などにおいては、スピーキング・テストが導入されている。それぞれは想定する英語使用場面・目的が少しずつ異なるため、画一的には言えないが、主にはあるトピックに対してスピーチを行うことや、ある事柄や出来事について説明すること、面接型テストで応答を行うことを通して、英語を用いて情報の発信能力を主に測定しているといえるだろう。しかし、面接型テストにおいて英語運用の際の態度が観点に含まれるものの、相手とのやりとりを成功させるための方略的な能力については反映できていないと考えられる。

このように、必要であるが測定が難しいやりとりに関わる能力・構成概念がある他にも、Folse はスピーキング能力の測定についての今後の課題として、以下のものを挙げている。まず、直接的にスピーキング能力を測定する場合、一人か少人数しか一度に評価することができず、時間がかかりすぎてしまうこと。また、評価に関しても、どの要素をどの比重で評価すればよいか明らかでない。そして、その要素もどのようにしたら測定できるだろうか？（例えば、発音の力を測る場合には、発音のエラーを数えれば良いのだろうか？）もう一つの方法として、間接的にスピーキング能力を測定する方法があるが、この場合、評価基準は明確になるが、本当に見たい力を測定しているかの妥当性の問題が生じる。

このような問題を一つ一つ解決していくことで、その他にも、別章で扱われているように、より包括的かつニーズにあったスピーキング能力の測定が可能となるだろう。

2. 実証研究の必要性

現在日本の英語教育において、L2でのスピーキング能力を評価する方法として広く用いられている形式は、スピーチと面接形式のテストである。しかし、現実世界では仕事における会議の場面や日常生活の友人との会話のような場面のように対等な相手がいる状態での協同的な会話構築が求められる場面多くあり、従来のテストでは幅広いコミュニケーション能力をカバーするためのスピーキング能力を測定しきれない可能性がある。そのため、現実世界と同様の *authenticity* を再現する方策として、L2でのペア/グループでのスピーキング・テストの必要性が考えられる。また、この型のテストは、*authentic* な文脈での協同的な会話構築にあたるコミュニケーション能力を測定できるだけでなく、教室活動でよりこうした活動の導入を促す波及効果があると考えられる。

そのため、本文ではL2学習者のための ペア/グループ・ディスカッションによるスピーキング・テストの探求として、ペア/グループ・ディスカッション・オーラル・テストについて、評価者の観点から、評価基準・構成概念・テスト実施と評価に影響する要因を調査した2つの実証研究論文に基づき、議論する。

3. 実証研究紹介

(1) Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26, 423-443.

■ Introduction & Background

・これまでは口頭での面接型テストが第二言語 (L2) のスピーキング能力を測定するために用いられてきた。

・しかし、1980年代から *peer-to-peer* でのインタラクションを含むペアあるいはグループでのスピーキング・テストが注目されてきている。

・ペア/グループでのスピーキング・テストは、面接型テストに比べて特有の文脈でない現実性を持ち、正の波及効果を持ち (Hildson, 1991)、複数の受験者を同時に評価できるという点で対時間・費用効果が良い (Morrison & Lee, 1985) という利点がある。

・Kramersch (1986) らの述べる *interactional competence* は、話し手がどのように *speech* をどう構成・配列するか、どのようにターン・テイキングを適用するか、どのように話し相手と会話を協同構築するか、によって定義される。

・面接型テストでは、面接官と受験者の間に立場の違いがあり、質問と応答に

より構成される一方、ペア/グループでのスピーキング・テストでは受験者同士での topic-building により高得点を得ることもある。

・これまで、ペア/グループでのスピーキング・テストは評価基準は評価のプロセスはあまり調査されてこなかった。

・データに基づく評価基準の作成方法として、これまで、学習者の speech や writing のサンプルを分析する方法が広く用いられてきたが、その他に評価者によるパフォーマンス評価の際の特徴を特定する方法がある。

■ 研究の目的

・ペアでのインタラクションの際に、評価者が何に注目しているか、を明らかにすること

・面接型テストと比べて、ペアでのインタラクションの構成概念は何か？

■ Method

・スペイン語を外国語として学ぶ 17 名の大学生（初級学習者）

・学習者はペアになり、10 分間「家族」「国民の休日」「友人」の 3 つのトピックについて、ディスカッションを続けた。

・12 名の評価者で初級学習者に教授の経験あり（11 名がスペイン語の native speaker）

・評価者はいくつかのディスカッションの映像を見て、評価の際の発話を行った。まず、映像を全体を見て、インタラクションの印象についてコメントを行い、次に映像を部分的に止めながら、詳細なコメントを行った。そのコメントのプロトコルが分析された。

・プロトコルは idea unit に分割され、それらは複数人で 5 つの分類: (a) body language, (b) listening, (c) turn taking, (d) topic cohesion, (e) anything other than interaction に分けられた（84%の評価者間信頼性）。不一致があった、turn taking と topic cohesion については、この 2 つの下位分類にもつ、interactional management が設定された。

■ Result

・評価者は以下の 3 つのインタラクションの parameter として、以下の 3 つの分類を用いていた。

(a) non-verbal interpersonal communication: gaze and body language

- ・評価者は「ペアが視線を合わせていることにより会話におそれを感じるものがなくなっている」という gaze があることを肯定するコメントを残している。
- ・また、body language に関しては、「body language により発話に感情を付加している」という肯定的なコメントと、「過剰な body language はことばによる表現能力の欠如」を示すという否定的なコメントの両者が存在する。

(b) interactive listening: supportive listening and comprehension

- ・評価者はインタラクションの際に、listening をする間の attention と engagement を示す方法として、supportive listening と comprehension 下位分類に持つ interactive listening についてコメント残している。

- ・supportive listening については、評価者の「あいづちにより相手の発話を続けることを促進している」というコメントに見られるように、back-channeling を示すことで聞き手としてもインタラクションに貢献する方法である。

(ただし、聴き手は理解をしていないことを隠すために、あうづちをうつことがあるため、注意が必要)

- ・comprehension については、インタラクションへの engagement を示すと共に、相手が発話に困った際に、言葉を補うなどの支援を含む。評価者は「一人が考えながら長く黙ってしまった時に、もう一方がことばや質問を返すことで、支援している」と評価するコメントを残している。

(c) interactional management: horizontal and vertical management

- ・評価者の「素早く答えることができなく、時間がかかってしまう」「turn taking に敏感になり、いつでも反応できるようにすることが大切」というコメントから、turn taking への注目があることがわか。るまた、「適切な質問を行うことで話していることに一貫性を持たせることができている」などのコメントから、topic を掘り下げる能力も評価している (horizontal management)。

- ・さらに、「一人がこれまでに話したことと関連づけることにより、トピック同士をつなげようとしている」というコメントは、topic 間の推移を行うことを評価することを示している (vertical management)。

■ Discussion

- ・上記の(a)~(c) はペアにおけるスピーキング・テストにおける、効果的なイン

タラクションの構成概念と適切な評価基準の発展に示唆を与えた。

- ・ペアの対話者は gaze や body language などの non-verbal interpersonal communication によってコミュニケーションを示す。
- ・ペアの対話者は、発話者への理解を提示、もしくは支援的なフィードバックにより、コミュニケーションを示す。
- ・ペアの対話者は、今続けているターンとトピックを超えたターンにより、コミュニケーションを示す。

(2) Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, 26, 161-186.

■ Introduction & Background

- ・グループでのディスカッション・スピーキング・テストは、同時に複数の受験者を評価でき、評価者も面接官としてのトレーニングがいないという点で、面接型テストに比した利点がある。
- ・またこのテストは、教師が現実世界でのコミュニケーション活動へより目を向けるという教室活動への波及効果をもつ。
- ・しかし、グループでのディスカッション・スピーキング・テストによる得点は、グループの参加者の性格に影響を受ける恐れがある (Bonk & Ockay, 2003)。
- ・本研究は、自己主張のある受験者と自己主張のない受験者の得点が、どの程度グループの参加者の自己主張のレベルによって影響を受けるか、を調査する。

■ Background

- ・Bachman (2001) の oral test performance のモデルにおいて、受験者の interactant がスピーキングパフォーマンスに影響を与えるとされている。
- ・Shohamy et al. (1986) では、グループ・オーラル・テストは評価の信頼性に困難を示し、他のスピーキングテストとの相関が.60 と高くなかった。
- ・Liski and Puntance (1983) は Talkative な性格を持つ参加者が発話のエラーが少ないことを報告している。
- ・Berry (2004) は、外向的な性格の参加者も内向的な性格の参加者も、外向的な性格のグループに入った時に、高いグループ・オーラル・テストの得点を得ることができた、ことを報告した。

■ 研究の目的

・自己主張のある受験者/自己主張のない受験者の得点が、グループの参加者の自己主張のレベル（自己主張の強い参加者のグループ・自己主張の強い/弱い参加者混合のグループ・自己主張の弱い参加者のグループ）によってどの程度影響を受けるか、を調査する。

■ Method

- ・協力者: EFL 学習者である日本人大学一年生 225 名
- ・評価者: スピーキング評価の訓練を受けた経験のある 34 人の Native speaker
- ・マテリアル:

(a) NEO-PI-R: 協力者の性格を測定するために用いた。neuroticism, extraversion, openness, conscientiousness, agreeableness の主な 5 観点を持ち、それぞれの観点はさらに下位の観点に分類される。自己主張性は、extraversion の一つであり、「ためらいなく発言し、グループのリーダーとなることを楽しむ性格」と定義される。

→結果から協力者を、自己主張性あり と 自己主張性なし に分類した。

(b) PhonePass SET-10: 日常生活の話される英語を理解し、ネイティブに近い英語の会話に反応を返す能力を計測した。54 の項目を持ち、read-aloud, repeat-sentence, short-answer question, sentence-build, open-question からなる。TOEFL などとも強い相関を示す oral proficiency テストである。

(c) group oral test: 4 人のグループで円になり、8 分間話し（議論し）続けた。グループは、(a) 全員自己主張性あり, (b) 自己主張性あり×3・自己主張性なし×1, (c) 自己主張性あり×1・自己主張性なし×3, (d) 全員自己主張性なしの四種類を設けた。評価者は円の外に座り、発音・流暢さ・文法・語彙・コミュニケーション方略のそれぞれの観点を 9 点満点で評価した。

・分析: PhonePass SET-10 の得点を共変数とし、自己主張性あり/なし それぞれの group oral test 得点を従属変数、グループ条件を独立変数とした MANCOVA を行った。

■ Result & Discussion

(1) 自己主張性ありの受験者 (N = 112)

- ・他のグループ参加者が自己主張性なしの場合、他の場合に比べ有意に高い得点を得た（全ての下位の観点において有意）。

- ・全員自己主張性ありのグループに入った時、他の場合に低い得点を得た。

（これらは PhonePass SET-10 に測定される個人の英会話能力を考慮している）

(2) 自己主張性なしの受験者 (N = 113)

- ・自己主張性なしの受験者の得点は、参加者の性格の異なるグループ間で有意な違いはなかった。

- ・(1)の結果の説明として、(a) 評価者が受験者のグループ・オーラル・テスト得点を、そのグループの参加者(の性格)の比較により評価してしまった (b) 受験者がそのグループの参加者の型に基づき、異なるパフォーマンスを示したの2つの可能性がある。(b) については、自己主張性のあり受験者が、他に自己主張のある参加者がいない時の方が発話の機会が多かった可能性を示唆する。

- ・また自己主張性なしの受験者にグループ型の差異がなかったことの説明として、Berry (2004) の、テストがハイスティクスである場合には、どの受験者もアグレッシブな参与を示す傾向があり、本研究のテストがハイスティクスな性質を持っていた、ということありうる。

- ・本研究の示唆として、評価者が受験者のグループ・オーラル・テスト得点を、そのグループの参加者内の比較により評価してしまった可能性から、グループ・オーラル・テスト評価者の訓練過程に、グループの参加者内の比較で得点を与えないように指導する必要性が提起される。

4. 考察 (2つの論文から)

まず、ペア/グループ・ディスカッション・オーラル・テストの構成概念として、受験者の gaze や body language などの non-verbal interpersonal communication, 発話者への理解を提示・支援的なフィードバック, 一つのトピックに対しターンを交換し続けることとトピックを超えたターンによるトピック内・トピック間の一貫性が重要であることが明らかにされた。これまでは主なスピーキング能力の構成概念として、発音や流暢さ、正確さ、内容妥当性など様々なものが入り入れられているが、この研究により受験者同士のペア/グループであるからこそ（面接型ではなく）含まれるべき構成概念があることが示

された。また、ペア/グループの型であるがために新たに影響を与える要因があると考えられる。具体的には受験者要因として、受験者の性格が挙げられる。今回の研究では、受験者の性格が評価者の評価に影響を与えたのか、実際のパフォーマンスに影響を与えたのかは区別することができないが、後者への影響も今後はさらに検討する必要性があると考えられる。

また、評価基準の作成についても、ペア/グループ・ディスカッション・オーラル・テストには、ターン・テイキングや **supportive listening** などが評価されるべきであることが明らかにされた。今後は、これらの評価対象がどのように尺度化され、尺度のそれぞれの段階にどのような **descriptor** が適切かを定める必要がある。

さらに、評価者が気をつけるべきこととして、ペア/グループとして複数人を同時に評価することは、対時間・費用効果はあるが、そのペア/グループの中で比較をして評価してしまうリスクを高める。実際に、評価者がグループ受験者の性格の影響を受けている可能性があり、与える得点が実際の能力を表すかについてはまだ猜疑心を残すところがある。そのため、評価者の訓練法についても工夫が求められる。実際のペア/グループ・ディスカッション・オーラル・テストの様子映像を見て、事前にグループ内の比較にならない姿勢を養うなどの方法が一つとしてあるだろう。

これらの今後の課題に加え、ペア/グループ・ディスカッション・オーラル・テストには、次のような問題も残される可能性がある。まず、このテストはたしかに **authentic** である性質はあるが、テスト的性格は避けられないことである。実際にハイステイク/ローステイクでパフォーマンス等が異なる可能性がある。また、ハイ・ステイクのテストとして用いられたとき、受験者同士は競争関係と協力関係の両面の性質を持つことになり、一般的な通常のディスカッションと異なる可能性がある。例えば、**Ducasse and Brown (2009)** で見られる受験者同士の **help** などの構成概念が影響を受けると考えられる。

最後に、教育現場への示唆として、以下の事柄が期待される。まずは、テスト機関によるペア/グループ・ディスカッション・オーラル・テストの導入により、複数の参加者による協同的な会話におけるスピーキング能力を測定できるようになることである。このことは、波及効果として、教師が教室へのペア/グループ・ディスカッション・オーラル・テストの導入へ目を向けることも期待

される。また、こうしたテストの実践と調査・改善を通して、現在教室で実際に行われているペア/グループ活動や教師の評価法へ示唆があることが期待される。こうして、一般的にインタラクティブなテスト・活動が普及することは、生徒のペア/グループ・ディスカッションへの姿勢と能力向上へとつながるだろう。