

オンラインスピーキングテストと面接式会話テストの 実践比較報告

ショー出口 香
カルビン大学

要 旨

コミュニケーション能力の習得は、外国語学習の重要な目標の一つである。しかし、言語四技能の中で、話すという技能は、限られた人材と時間で適切な評価を行うのが困難なコミュニケーション能力である。本稿では、話す技能を評価する新しい試みとして実施したオンラインでのスピーキングテストと、従来の面接式会話テストとを比較して報告する。オンラインテストでは、Web 上で、学生は、教師が準備した短いビデオクリップを視聴し、指示に従って自分の反応を録音した。面接式会話テストでは、試験官と学生が対面でロールプレイを行った。そして、この二種類のテストの妥当性、公平性、緊張度に対する学生の認識について、アンケートを用いて調査した結果、各テストで評価される能力や評価の公平さに差が見られた。よって、今後は、それぞれのテストが評価対象とする技能をより明確に定義づけ、相互補完的にカリキュラムに組み入れていくことが望ましい。

1. はじめに

大学レベルでの初・中級外国語コースの多くが、言語の四技能の総合的育成を目標として掲げていると考えられる。もちろん、この四技能の中でどれを優先するかは学習者によって異なるであろうが、「話せるようになりたい」という思いは、程度の差こそあれ、大学レベルで初・中級外国語コースを履修している者に共通するものである (Rivera & Matsuzawa 2007)。しかし、この話すという技能に対し、初・中級レベルの外国語コースで、実際にどのように教授、学習、評価が行われているのかというこ

とには、あまり目が向けられていないようである。もちろん、これらの四技能は個別に学習／習得されていくものではない。それでも、四技能の中でどの技能の習得がどの程度進んだのか、または遅れているのか、主観的な学習者個人の感覚だけでなく、そのコースにおける客観的な成績として学習者に提示することができれば、学習意欲や目的意識の維持に効果があるのではないかと思われる。

それでは、話す技能は、初・中級レベル外国語コースの成績の中に、どのように反映されているのだろうか。文型や語彙の知識、またその知識を使用して読んだり書いたりする能力に関しては、どの言語コースにおいても、段階に応じてクイズやテストなどを定期的に行っているのが一般的であろう。それに対し、話す技能を定期的、継続的に評価するクイズやテストは、同様に行われているであろうか。また、聞く・話す技能が評価の対象となっているテストやクイズは、読む・書く技能対象のものと同等の比重を成績の中で占めているだろうか。これらの疑問に対する答えは、大学でのほとんどの初・中級レベル外国語コースで否だ、と予測される。なぜなら、話す技能は、限られた時間と人材で適切な評価を行うのが難しいという認識が一般的だからである。そこで、本稿では、まず、なぜ話す技能の評価が困難だと認識されているのかを考察する。そして、話す技能を学習段階に従って定期的に評価することを目標とし、また成績全体に占める話す技能の評価比重を大きくするための試みとして実施した、オンラインでのスピーキングテストについて、面接式会話テストと比較しつつ、報告する。

2. 先行研究

2.1. 話す技能とは

まず、本稿で論じる話す技能の定義を明確にしておきたい。実際に初・中級レベルのクラスで指導が行われている話す技能とは何か、Bygate

(1987) が提案した *model of speech as a process* に基づいて以下に示す。話すということは、第一に、状況判断に基づいて何を話すかを考えること (*planning*) に始まる。話す内容が決まったら、第二に、それをどう表現するかを決める (*selection*)。この段階で、語彙、文法、文化に関する知識を活用する。第三に、その表現を実際に口に出して、相手に理解できるように産出 (発話) する (*production*)。さらに、会話とは相手がいて成り立つものと考え、相手の発話を理解しそれに即時に応じて、適切で建設的な発話をする (*reciprocity*) が重要な要素と考えられる。この、柔軟に建設的に対話を継続する技能は、コミュニケーションを重視する外国語教育においては、初・中級レベルであっても指導するのが理想である。しかし、この技能が、評価を含めたあらゆる面で他の技能と同等に扱われている外国語のカリキュラムは、まだ多くはないのが現実であろう。上記の三段階から形成される話す技能を評価する場合、最終段階で産出された *production* を対象とするのであるが、「書かれたもの」に比べて「話されたもの」を定期的に評価することは、初・中級レベルにおいてはあまりなされていないと思われる。次の節では、その理由について、これまでに報告されているスピーキングアセスメントの内容を再検討することにより、明らかにしたい。

2.2 既存のスピーキングアセスメント利用の問題点

2.2.1 話す能力検定のアセスメント

これまでに詳細に報告が行われてきた話す技能のアセスメントは、学習者が対象言語を用いて、どの程度、母語話者に近いコミュニケーションを行うことができるかを評価することを目的とした、*standardized proficiency* に基づいて評価するものが多い。その代表的なものとして、OPI, SOPI, COPI, The CoSA, SPOT, JSKIT, COETなどが挙げられる。これらのアセスメントを、特定の初・中級コースの成績評価の一部として使用するのはい

題である。その理由の多くは、Hirai & Koizumi (2009) が指摘している、話す技能を評価するアセスメント開発の困難点と重複する。第一に、評価の対象となっている能力 (constructs) が、これらのアセスメントと初・中級コースで必要なテストとは大きく異なる。初・中級コースで必要なのは、既習の知識 (文法や語彙) を用いて、意味のある発話をする事ができるかの評価 (achievement test) である。Bygate (1987) のモデルに基づくと、初・中級コースで行われる話すテストの場合、何を言うか (planning) とどの文構造を使うか (selection) はすでにある程度設定されている。評価の対象は、学習者が既習の知識を使用して、与えられた状況に合わせて、ある表現を作り変え、それを相手に理解できるように産出すること (production) ができるかだと考えられる。第二に、OPI のようなアセスメントは、特定のトレーニングと豊富な経験を積んだ試験官/採点者が学習者と一対一で行う必要があり、また、そのアセスメントを開発した機関に料金を支払い実施の許可を得なければならないなど、頻繁に利用するのが困難だという点 (practicality) が挙げられる。第三に、OPI などのアセスメントは、一定の教材を念頭において作成されたものではないため、ある特定コースの授業内容とは関連性が低い。よって、アセスメントそのものに、その後のコース内の指導と学習への建設的な効果 (impact on teaching and learning) があるとはあまり期待できない。以上の点から、proficiency を評価するために開発されたアセスメントを、成績の一部として定期的に使用している初・中級外国語コースは少ないと考えられる。そこで、次に、特定コースの評価にも応用可能な新しいスピーキングテストについて検討したい。

2.2.2 Direct と Semi-direct のスピーキングテスト

話す技能の評価方法は、試験官が一対一で学習者と対話して評価対象となる発話を産出させる direct test と、試験官の代わりに、テープ、コンピ

ューター、紙面で何らかの情報（発話のきっかけとなるもの）と指示を提示して発話を産出させ、それを録音して評価を行う **semi-direct test** に大きく分けられる。**semi-direct test** では、現実の会話相手不在で発話が行われるが、そのタスクには多様なものがある。

Weir & Wu (2006) は、短い文章を読み上げる (**read aloud**)、絵や写真を描写する (**picture description**)、質疑応答 (**question and answer**) というタスクを、中級レベルの英語学習者に使用した。また、Hirai & Koizumi (2009) は、やはり中級レベルの英語学習者を対象に、**Story Retelling Speaking Test** を開発、実践し、学習者がこのテストについてどう感じたかを報告している。このテストでは、学習者は、短い話を読み、内容確認の質問に答えた後、その内容をできるだけ詳細に思い出して話し、さらに、話に対する自分の考えを述べるという課題を行った。そして、学習者はこのテストにおいて、読解能力よりもむしろ話す能力の方を評価されたと認識した、と結論づけている。この二つの研究で使用された課題は、クラスで学習したトピックや文型／語彙と関連を付けやすいため、初・中級コースで定期的に行う **achievement test** にも応用できるであろう。ただ、これらの研究では、聞き手が存在せず一方的に発話を行うことについて学習者がどのように感じたかは、調査されていない。

一般に、学習者が話す能力について言及するときに思い浮かべるのは、「誰かと話してコミュニケーションをとる」対話であろう。録音装置やコンピューターという架空の相手に発話を行って、本当に自分の話す力が評価されているとは感じられない、つまり妥当な評価方法だとは思えない、という認識を多くの学習者が持つのではないかという危惧を、教師が抱くかもしれない。この点について、Qian (2009) が、**direct** と **semi-direct** のスピーキングテストを学習者に両方行わせ、その反応と視点を比較し、**semi-direct test** に対して批判的な態度は多くなく、むしろ、好意的でも反好意的でもないという中立的な態度が 58% を占めていた、と報告してい

る。ということは、今後、コンピューターを用いた新しいスピーキングテストが開発実施された場合、最初から懐疑的な態度で臨む外国語学習者は多くはない、と予測していいのではないか。Qian (2009) が使用したタスクの中に、就職する際の面接試験という設定での質疑応答が含まれている。これは、*direct test* でもよく使用される設定で、このような具体的な状況における会話を、実際の相手不在で行ったテストにおいても、学習者の多くが中立的な態度を示したということは、興味深い。

以上、ここで考察した新しいスピーキングテストはどれも、コースにおける各段階での達成度を確認する目的で行われたのではなく、学習者の能力がその時点でどのレベルにあるのかを判断する *proficiency test* として作成／実施されたものである。しかし、OPI 等に比べて、時間、人材、機材などの実用性 (*practicality*) の面では、これらの *semi-direct test* の方が、外国語コースで定期的に行う *achievement test* として適していると考えられる。しかし、Qian の研究では、*direct* と *semi-direct* どちらのテストにも複数の異なるタイプのタスクが含まれていたため、初・中級コースで最も一般的な一対一の面接形式のスピーキングテストと同じような状況を、オンラインで機械を相手に行った場合、学習者の反応や認識がどう異なるかは明らかではない。そこで本研究では、この点に焦点を当てて、*direct* と *semi-direct* のスピーキングテストの比較検討を報告する。

2.3 スピーキングテスト開発／比較の基準

テストを作成するとき、さらには、作成したテストを比較し考察を行うとき、ポイントとなる項目を明確にしておく必要がある。本稿では、Luoma (2004) が考案した *modular specifications* を参考に、面接式会話テストとオンラインのスピーキングテストの違いを考察したい。

Luoma (2004) は、スピーキングテスト開発を、四つの異なる側面に分けてとらえ、それぞれの側面における項目の変化が互いに影響を与え合いつ

つ、理想的なテストが作成されていくべきだとしている。その四つの側面とは、**construct specifications**, **task specifications**, **assessment specifications**, **history file** である。本稿で考察するオンラインテストは、新しく作成、実施したもので、参照すべき以前のテスト結果がないので、**history file** については割愛する。**construct specifications** とは、テストを受ける学習者は誰か、目的は何か、どんなスピーキング技能が評価の対象となっているのか、その技能はどのような理論に基づいて抽出されたのか、などを明確に記述したものである。この記述が、テスト作成の基盤となる。**task specifications** とは、どのようなタスクを通じて対象となるスピーキング技能を学習者から引き出すかを、明確に記述したものである。また、テストを行う試験官が、事前にどんな指示を与えるのか、また試験中にはどんな役割を果たすのか、などを記述したものも、ここに含まれる。**assessment specifications** とは、対象となる技能を、どのような基準に従って、いくつのレベルで評価を行うのかを、詳細に記述したものである。この記述に基づくことによって、採点者は公平な評価を下すことが可能になる。テスト作成／開発の中で、評価の対象となる技能が変われば、その技能を引き出すタスクも変える必要がある。ある特定の技能が、複数のタスクによって統合的に引き出される可能性もある。タスクが変われば、たとえ目的の技能は同じであっても、評価基準はタスクに合わせて変える必要が生じる。このように、テスト作成／開発のプロセスは、この三種類の記述が整合性を持つように進められていくのが理想であり、それがこのモデルの主眼である。本稿では、このモデルを基に、二つのスピーキングテストを比較する。

3. 使用したオンラインシステムの概要

本稿で報告する **semi-direct** のスピーキングテストは、外国語学習における口頭練習／評価をサポートする目的で開発された、**Speak Everywhere**

(以下「SE」と略称)というオンラインシステムを利用して作成／実施した。このシステムをカリキュラムに組み込んだ日本語コースの実践報告を、池田・深田 (2012) が行っている。このシステムでは、まず教師が、短いビデオクリップと、学生がそれを見て行うべき活動の指示を準備する。学生は、インターネット上でビデオクリップを視聴し、指示に従って発話を行う。その発話は、システムによって録音される。学生が、録音された自分の発話を何度でも聞き返せるよう設定でき、また、タスク内の単一の発話もタスク全体も一定期間何度でも繰り返せるよう設定することが可能である。学生の発話は、その都度、同一の音声ファイルに上書き保存される。教師は、学生が最後に行った発話の音声ファイルを聞いて、評価を行う。

4. 実践報告

4.1 コースの概要

本研究は、アメリカの州立大学日本語プログラムの四学期目 (202) のコース (三学期間で合計約 225 時間すでに受講) で、2010 年春学期に行われた。当コースでは、『なかま 2』 (Hatasa 2000) の 7 課から 10 課までを学習した。在籍した 79 名の学生の中で、SE を使用するのが初めてだった学生は 10 名のみで、残りの学生は、2009 年秋学期に引き続き二学期目であった。よって、大部分の学生がこのシステムの使用に慣れていたと考えられる。コース最終成績 100% の中で、SE を用いた宿題が 6%、SE 上でのテストが 4%、面接形式の会話テストが 3%、学期末のプレゼンテーション時の発音と流暢さが 2% をそれぞれ占めていた。つまり、話す技能関連の評価は成績全体の 15% であった。

本稿で比較を試みるオンラインのスピーキングテスト (semi-direct) と面接式会話テスト (direct) の共通項目 (shared construct specifications) は、日本語 202 の学生を対象に、学習者が特定の既習項目を口頭で状況に応じて使用できるかどうかを評価し、またどの点がどれだけ達成できているか、

できていないかのフィードバックを返すという目的があり、コースの最終成績の一部となる、という点である。

4.2 面接式会話テストの概要

このテストは、2010年春学期の第八週目に、授業時間内に通常の教室で実施した。学生一人につき約10分を費やした。

評価対象は、待遇表現の知識を活用して意味のある質問を行い、会話を続ける技能である。また、会話とは、一方向的な発話だけでなく、相手と共同して情報や意見の交換を行い構築するものだ、という考えに基づき、学習者がいかに会話に積極的に参加しようとしているかも評価しようと試みた。つまり、ただ単に準備してきた質問を試験官に尋ねるだけでなく、試験官の答えに対して適切なフォローアップの質問またはコメントを、自然な会話の流れを保持しつつタイムリーに返せる技能も、評価対象とした。

このテストのタスクは、試験官と学生との一対一のロールプレイである。設定は、日本から大学の研究者が、講演するためにアメリカの大学を訪れたので、学生は、大学新聞に記事を書く目的で、この研究者に日本語で簡単なインタビューを行う、というものだ。会話の流れは、まず、学生は簡単な自己紹介をし、またインタビューの目的を相手の研究者に説明する。そして、事前に準備した八つの質問を尋ねる。質問は、研究者の学歴、職歴、家族構成や趣味など個人的なこと、そして、今回のアメリカ訪問の予定の四つのカテゴリーについて、二つずつ聞くように指示した。質問をする順序は、学生に自由に決めさせた。さらに、より自然な会話を学生に行わせるため、試験官の答えに対して、できる限りフォローアップの質問かコメントをするように指示した。このテストのタスク内容と評価基準は、テスト前に授業時間一コマ（50分）を使って詳しく説明し、また、学生はペアで練習を行った。テストの際、英語での簡単なメモを見ることを許可した。テストは、授業時間内に通常の教室で、各学生10分以内で一人

ずつ行われた。試験官は、そのセクションを担当していないティーチングアシスタントと筆者が分担した。試験官が、ロールプレイの相手役となり学生からの質問に答えた。同時に、採点も各試験官がその場で行った。

このテストの評価は、特定の情報（質問内容）ごとに、新しく学習した項目を、適切にまた正確に使用して質問する意図が見られたかを基準とした。例えば、相手の学歴に関する質問では、「大学で何を勉強なさいましたか」という発話に満点の 0.5 点が、「大学で何を勉強なさいましたか」という発話に 0.4 点、「大学で何を勉強いたしましたか」という発話には 0.3 点、「大学で何を勉強しましたか」という発話には 0.2 点が、それぞれ与えられた。つまり、待遇表現使用を試みて間違いがある発話に、文法面では正確だが待遇表現を全く使用していない発話よりも高い点を与えた。加えて、相手の答えに対して、意味があって建設的なこと、つまり会話をさらに展開させる意図のはっきりとしたフォローアップの質問かコメントができたかどうかを評価した。このフォローアップの発話においては、文法上の正確さを一切考慮しなかった。たとえ文法的に不正確であっても、意図がはっきりと示されていて、理解可能な発話であれば、一つの完全な発話と見なした。この項目の満点の 4 点は、8 回以上建設的で異なるフォローアップの質問かコメントを行った場合に与えられた。

4.3 オンラインのスピーキングテストの概要

オンラインのスピーキングテストは、2010 年春学期の第 2 週目と第 14 週目の二回行った。テストの目的は、既習文法項目を使って自分が伝えたい内容を表現することができるかを評価することと、学生の発話に教師からのフィードバックを与えることであった。テスト内容と評価基準について、コンピューターラボで 50 分（一クラス）かけて説明し、タスク練習を行った。テスト期間は一週間で、その間、学生は、各自好きな場所で好きなだけ時間をかけてテストを行った。教科書や自分のノートを見ながら

タスクを行う、また、友人の助けなどに頼る、といった外的要素は全くコントロールしなかった。

評価対象は、自分がその時点で持っている語彙／文法知識を利用して、自分自身と自分の経験や考えについて、伝えたい内容を明確に口頭で表現することができるか、であった。一回目のテストは、主に、前学期の 201 で学習した語彙／文法の知識、二回目のテストは 202 で学習した語彙／文法知識の適用に焦点を当てた。試験問題（録画された質問や教師の自己紹介）を学生は何度でも視聴できるという点で、自然な会話とは異なる。しかし、ビデオクリップで提示された教師の発話を出発点として、いかに話を発展させられるか、いかに多くの情報を提供できるかを評価する意図で作成した。

このテストは二部で構成された。第一部では、学生は、ビデオ上の教師が尋ねる 10 の質問に答えた。既習の文構造や語彙を使ってできるだけたくさん話すよう、また、録音された自分の回答を聞き直し、満足するまで繰り返すように奨励した。質問内容は、主に、学生個人に関すること、日本語学習に関することであった。第二部は、教師が簡単な自己紹介をしているビデオクリップを視聴し、その内容に関して質問を作成／録音させた。質問内容は、家族、学歴、仕事、趣味、夏休みの予定の 5 つのトピックにつき、一つずつ作成するよう指示した。

図1 オンラインテストの第二部の例

Course: CC_JPN_Test Exercise: Online Test 2 Part 2 Asking Questions Item 1 of 5

**Speaking Test 2
Part 2: Ask Questions**

Instruction:
先生の自己紹介を聞いてから、この先生に、5つ質問をしてください。

The teacher in the video will talk about herself. Take notes while listening. Listen to the questions that you made in the beginning of this semester. You need to use **new structures** and more **formal and polite** expressions you learned in JPNS202.

(1) Ask a question about her family.

REDO SUBMIT & NEXT
LISTEN S<=>T T<=>S
STOP ?

Playback ended

このテストの評価は、第一部も第二部も、学生の音声ファイル一つ（まとまりのある一つの発話）につき、コミュニケーションをふまえた情報内容、文構造、文法の正確さ、発音の正確さ、流暢さの五項目についてそれぞれ五段階で行った。第一部の質問に対する回答では、会話をさらに続け、たくさんの情報を提供しようという意思が表れている回答に最高点を与えた。文構造では、201 で学習した構造を使用している回答に最高点を、101/102 で学習した、例えば「名詞＋です」「名詞＋助詞＋動詞」のみの単純な構造の文を使用している回答には、一レベル低い点を与えた。採点は、各セクションを担当している教師が、自分のセクションに在籍している学生に対して行った。採点を始める前に、教師全員で、録音された学生の回答をいくつか試聴して話し合い、採点基準について合意を得られるよう努めた。

4.4 学習者からのフィードバック収集

面接式会話テスト終了後の第9週に、オンラインで自主参加のアンケートを実施し、47名の学生から、この二つのスピーキングテストに関してフィードバックを収集した。テスト時にどのくらい緊張したか、テストは公平だと感じたか、どのくらい難しいと感じたか、好意的に感じたか、発音/流暢さ、コミュニケーション能力、そして、既習の構造を応用できる能力がどのくらい評価されていると思ったか等を、それぞれ10段階で評価してもらった。表1にその結果をまとめた。

表1 学生によるテストに対する十段階評価

	オンライン		面接式		t値
	平均	標準偏差値	平均	標準偏差値	
緊張度 (n=50)	3.56	2.64	7.28	1.88	1.153*
公正さ (n=47)	7.83	2.32	8.30	2.22	0.128
易しさ (n=47)	5.96	2.02	5.34	2.36	0.152
好意度 (n=47)	6.30	2.43	6.60	2.41	0.476
流暢さ/発音 (n=47)	6.49	2.11	7.43	2.10	0.014
コミュニケーション能力(n=47)	5.85	2.13	7.32	2.18	0.001
言語知識の活用 (n=47)	7.19	2.22	7.34	1.93	0.604

(*p値<0.3)

5. 考察

5.1 評価の対象となっている言語能力 (constructs)

まず、この2つのテストで評価の対象となる言語能力について考察したい。表1にあるように、学生に「日本語を使ってコミュニケーションする能力を、このテストでどのくらい発揮できたと思うか」を十段階で評価してもらった結果、面接式の方がやや平均値が高かった。面接式会話テストは、予め会話内容は定められていたが、対話相手の発話に対して即時的に適切な反応を返し会話の発展に積極的に参加する、というコミュニケーション能力の一側面を評価することを目的としていた。この能力は、準備して来た質問をするだけでなく、質問に対する相手の答えに応じて、何らかのコメントまたは新たな質問をして会話を発展させるよう学習者に指示す

ることによって、引き出された。このテストを通じて会話能力が発揮できたと認識している学生のコメントには、次のようなものがある。

- It was good that we had to make out own responses since it made it more like a real conversation.
- If you wrote down all the responses and just memorized it, then it is just repeating sentences rather than really speaking the language. Though the comments part might help us to really practice our listening and speaking skills.

これに対し、オンラインテストは、何度でも相手の発話を聞き返し、またそれに対する反応も何度でもリハーサルすることができるという点で、自然な対話の状況とは言えない。この点は、学生からも、コミュニケーション能力がオンラインテストに反映されていないと感じられる理由として以下のように指摘された。

- The final responds in the online test are the results of many previous practices. It is not your immediate respond, therefore not a true reflection of your speaking ability.

これらのコメントから、オンラインテストは、対話相手に対して適切な反応ができるという能力を評価するにとどまっていると学習者は考えており、面接式テストは、加えて、その反応を即時的に実行できるかという能力も、同時に評価していると考えられている、と言えよう。

「既習の文法／語彙を適切に使用する能力をどのくらい発揮できたと思うか」という質問に対する学生の十段階評価の平均値は、予想外に、ほとんど差がなかった。オンラインテストは、教科書やノート、または誰かの助けを得るなど、学生個人の能力以外に頼る可能性を制限しなかった。よって、教師の視点からは、厳密に、学生の既習言語情報の使用能力を評価しているとはいいがたい。しかし、学生は、どちらのテストにも、この能力が同じ程度反映されていたと認識していたようである。同様に、「流暢さ／発音」の能力に関しても平均値に大きな差はなかった。

まとめると、今回実践した二種類のテスト間で、評価の対象となる言語能力の違いは、即時的、建設的に会話を展開させていくというコミュニケーション能力の一要素のみと言えよう。よって、可能な言語活動が限られている初・中級レベルでは、SE を利用した semi-direct のスピーキングテストの妥当性は面接式テストと大差ないと考えられる。

5.2 学生の不安感 (anxiety level)

面接式テストにおける学生の不安感が高いこと (Shomoosi, Kassaian, & Ketabi 2009)、学生の性格 (緊張しやすいかどうか) がテストでのパフォーマンスに大きな影響を与えることは、外国語教師としての体験から否定できない。よって、今回実践したオンラインテストは、学生の不安感軽減と授業時間維持のため、あえてテスト実施の場所と時間を設定せず、学生の意思にまかせた (Kessler 2010)。また、質問を聞くことも回答することも、何度でも繰り返せるように作成した。この「緊張度を下げる」というオンラインテストの目的が、学生からも支持されたことは、アンケート結果が示している。「テストを受けている時どのくらい緊張したか」に対する十段階評価の平均値は、オンラインテストの方がかなり低く、学生の緊張度が低かったことは明らかである。その理由は、次のコメントに集約されている。

- The online test was not face to face. I can write down the sentences first and record later. Also, it was possible to listen to the record again and again.

即ち、実際の会話相手がいなく、発話内容を準備できたこと、そして繰り返しが可能だった点である。面接式テストで緊張した理由としては、会話相手が存在するだけでなく、その相手に同時に評価もされていること、チャンスが一度きりだという点が、学生から挙げられた。

学生のコメントの中には、実際に相手がいって発話することを好意的にとらえているものと、全く逆に否定的にとらえているものがあった。学習

者の性格によって、テストに臨む態度、そしてテスト中の行動に差が生じたことが伺える。否定的なコメントには、以下のようなものがあった。

- Face-to-face interview is always more nervous because I can see the interviewer's face expression while I'm answering.
- I always feel very stupid speaking Japanese to native Japanese speakers. This is true even if I didn't say anything wrong. I also worry about them asking follow-up questions and not thinking fast enough to come up with a response.

反対に、実際の話し相手がいることによって、むしろリラックスできたという以下のようなコメントもあった。

- I like this one better even though I may get nervous. The sensei was understanding and helpful, so it relaxed me in some ways, but being in front of someone makes it harder to think on your toes.

また、相手とのインターアクションを通じてコミュニケーションの修復を図るというストラテジーが運用できた学生にとっては、面接式の会話でもあまり緊張しなかったようである。

- I liked that I could ask my interviewer to repeat herself or to say it in different words.

テスト実施時の学生の不安感／緊張度の軽減という面においては、オンラインテストの方が効果的であったと言えよう。それに比べて、面接式テストの方が学生の緊張度が高まる傾向が見られたが、その度合いは学習者間の個人差が大きいようである。

5.3 評価の客観性と公平さ (objectivity / fairness of evaluation)

対話技能を評価する際の問題点の一つは、客観性の維持にある (Luoma, 2004)。どんなに詳細で明確な評価基準が記述されていたとしても、各試験官が、会話の相手を務めつつ評価も同時に行う場合、評価者間に差が生じるのは避けがたいことである。オンラインテストでは、評価をテストと同時に行うのではなく、録音された発話を聞いた後に行った。評価が難し

い特定の発話や複数の学生の発話を、評価者全員で聞き、話し合っ合意を得ることができた。つまり、評価者間の差を軽減し、より客観的な評価を下して信頼性を維持することが可能だと考えられた。この点を、学生はどう認識したのだろうか。アンケート調査では、「どのくらいこのテストをフェア（公平）だと思うか」に対する十段階評価の平均値は、予想に反して面接式テストの方が多少高い傾向にある。その理由として、オンラインテストに関しては、何度でも繰り返せる点を好意的にとらえるか否定的にとらえるかによって、テストそのものを公平だと見なすかどうか意見が分かれたようだ。繰り返せるから公平だという認識と、繰り返せるからこそ公平ではないという相対する認識が、コメントの中に見られた。面接式テストに関してはやはり、以下に代表されるように、試験官の間で採点に関して個人差があるのではと危惧している声が多かった。

- Since it is graded by different instructors, it is so unfair that some instructor gives a good grade and someone grades very picky.

このような、試験官の個人差に関する危惧を示唆するコメントは、オンラインテストに関しては、同じように特定の教師によって採点が行われたにも関わらず、全く見られなかった。学生のコメントからは、面接式会話テストの方が、評価の客観性について学生の不信感が高い傾向が窺えた。

5.4 実施の容易さ (practicality)

面接式会話テストの問題点の一つは、実施に時間のかかることである。学生数の多いコースで、限られた授業時間内に一对一の会話テストにどのくらいの時間を当てるかは、カリキュラムを構築する担当者の頭を悩ませる。時間をかければ、対話能力をよりの確に多岐に渡って評価することができると考えられるが、同時に、それだけ授業時間が削られることにもなる。SEを使ったスピーキングテストは、コンピューターラボで同時に行うことも、また、学生にある一定の期間を与えて、その期間内に好きな場

所で行わせることも可能である。貴重な授業時間をテストに当てる必要が無くなる。

5.5 テストが生み出す学習効果 (impact on teaching and learning)

本稿のオンラインテストの特徴の一つは、学生が何度でも繰り返せるという点である。どのような「繰り返し活動」を、何人の学生が何回行ったか、学生が自己申請した結果を表2にまとめた。

表2 繰り返しを行った回数とその学生数

	質問を聞き直す	録音をし直す	自分の答えを聞き直す
5回以上	10人	6人	3人
3、4回	26人	4人	5人
2回	11人	13人	12人
1回	0	23人	22人
ゼロ	0	1人	5人
合計	47人	47人	47人

テスト質問を聞き直すこと、自分の回答を録音し直すことは、回答者47人全員が行っていた。録音した自分の答えを聞き直すことは、5名以外の学生が、最低一度は行っていた。ここから、学生は、質問を聞き直し、答えを言い直し、また、自分の答えを聞き直すことを、各自必要に応じて繰り返して行っていたことが明らかだ。教室での限られた授業時間内で、このような繰り返しを行うのは困難である。また、学生によって、何を、どのくらい繰り返す必要があるか異なるであろう。SEを使用することによって、「繰り返し活動」の個人化が実現されたと言えるだろう。

しかし、これらの「繰り返し活動」により、実際にどのような学習が行われたのかは、今回の学生からの回答では明らかではない。以下のようなコメントから、既習の語彙や文型、文法規則を意識的に利用して意味のある発話を行うこと、また、自分の発話の間違いに気づきそれを自ら訂正することが学習されたと推測できる。

- You could fix problems and think of the answers.

- It was more difficult and made us use more sentence structures while still allowing us more chances at getting the questions right and allowing us all the time we needed.
- I can practice many times. Also it helps me in applying the grammar and new vocab by preparing responses myself.

上記のコメントから、オンラインテスト中の「繰り返し活動」が、二次的学習効果をもたらしたと考えられよう。この学習効果は、面接式会話テストでは得られないものである。

6. まとめと課題

本稿では、面接式会話テストとオンラインでのスピーキングテストを実施し、学生のテストに対する認識を比較考察した。話すという技能のうち、状況に応じて、何を話すか（planning）、それをどう表現するか（selection）、実際に言語として産出する（production）という面においては、どちらのテストでも評価がなされたと学生は認識していた。しかし、対話相手の発話に対して即時的に反応して会話に積極的に参与する技能（reciprocity）は、オンラインテストでは評価が難しいものだと学生は考えている、と言えよう。よって、今後、初・中級レベルでの話す技能の評価を定期的／継続的に行うためには、二種のテストを相互補完的に作成／実施していくことが効率的かつ理想的である。面接式会話テストでは、対話を進める上での reciprocity に関する能力に焦点を絞り、オンラインテストでは、既習言語情報を状況に応じて口頭使用する能力のみを対象として、それぞれアクティビティと採点基準を作成する必要があるだろう。何よりも大切なのは、学習者に、各テストの目的（評価の対象となる技能）と評価基準を、より明確に提示することであろう。

本研究では、オンラインテストの公平さに対する学生の評価が高くなかった。これは、繰り返しの回数、テスト時間と実施場所を制限することで改善できると考えられる。しかし、その場合、学生の緊張度が高まること

が予測される。公平性の維持と緊張度の軽減を、同時にいかに図るかが課題の一つである。

オンラインテストでは、二次的教育／学習効果を期待できることが学生のコメントから推察された。よって、SE を用いたアクティビティーを、クイズまたは教室外課題として定期的にカリキュラムの中に組み入れることが勧められる。既存のワークブックと平行して、同じ言語知識を口頭で使用／練習する課題を作成するのである。今後、話す技能をさらに細分化し、各技能をより明確に学習者に提示し、異なる技能を対象とするテスト、クイズ、教室外課題を、SE を活用してカリキュラムの中に組み入れていくことを目指していきたい。

注

- * 本稿は、2011年8月7日にカナダ日本語教育振興会の年次大会にて行った口頭発表「Online Speaking Test と従来の面接式会話テストの比較：評価対象は、文法能力かコミュニケーション能力か」の内容に補足・修正を加えたものである。執筆の過程で査読者の方からいただいた貴重なご意見やアドバイスに、ここで改めて感謝の念を示したい。

参考文献

- 池田順子・深田敦 (2012) 「Speak Everywhere を統合したスピーキング重視のコース設計と実践」『日本語教育』152, 46-60 日本語教育学会
- Bygate, Martin. (1987). *Speaking*. Oxford, UK: Oxford University Press.
- Hatasa, Yukiko., Hatasa, Kazumi., & Makino, Seiichi. (2000). *Nakama 2: Japanese Communication, Culture, and Context*. Boston: Houghton Mifflin Company.
- Hirai, Akiyo. & Koizumi, Rie. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly* 6, 151-167.

- Kessler, Greg. (2010). Fluency and anxiety in self-access speaking tasks: the influence of environment. *Computer Assisted Language Learning* 23(4), 361-375.
- Luoma, Sari. (2004). *Assessing Speaking*. Cambridge, UK: Cambridge University Press.
- Qian, David D. (2009). Comparing direct and semi-direct modes for speaking assessments: Affective effects on test takers. *Language Assessment Quarterly* 6, 113-125.
- Rivera, Grady M. & Matsuzawa, Chie. (2007). Multiple-language program assessment: Learners' perspectives on first- and second-year college foreign language programs and their implications for program improvement. *Foreign Language Annals* 40(4), 569-583.
- Shomoosi, Nematullah., Kassaian Zohreh., & Ketabi, Saeed. (2009). Variation of test anxiety over listening and speaking test performance. *Iranian Journal of Language Studies* 3(1), 65-78.
- Weir, Cyril J. & Wu, Jessica R. W. (2006). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing* 23(2), 167-197.