Kobe University Repository : Kernel

PDF issue: 2025-07-11

ダイアローグ型タスクによるスピーキング能力評価研究の動向 ： 学習者対話型テストを中心に

小泉，利恵

(Citation)
Learner Corpus Studies in Asia and the World,3:27-42

(Issue Date)
2018-03-12

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81010116

(URL)
https://hdl.handle.net/20.500.14094/81010116

ダイアローグ型タスクによるスピーキング能力評価研究の動向
—学習者対話型テストを中心に—

小泉　利恵（順天堂大学）
rkoizumi@juntendo.ac.jp

Recent Studies on Speaking Assessment Using Paired and Group Oral Tasks

KOIZUMI　Rie (Juntendo University)

概要
面接者と受験者が対話を行う面接型のスピーキングテストは，1980 年代後半にその妥当性の低さが指摘され，その後学習者同士が対話を行うテスト形式に注目が集まるようになった。学習者対話型テストには，他形式では得られない利点と，避けられない弱点がある。本稿では，学習者対話型テストの最大の利点と思われる，多様な発話機能を引き出しやすい等の構成概念の特徴について特にまとめ，学習者対話型テストとそのコーパス開発の可能性について論じる。

キーワード
ペア型口頭テスト，グループ型口頭テスト，構成概念

1. はじめに

スピーキング能力を包括的に測るためには，モノローグ（発表）だけでなくダイアローグ（やり取り，対話）を引き出す必要がある。2020～2022 年に全面実施となる小学校・中学校・高等学校学習指導要領の外国語（英語）においては，「話すこと」が「発表」と「やり取り」の 2 つに分けられ，「聞くこと」「読むこと」「書くこと」と合わせて 5 領域として指導することになっている。

ダイアローグは，面接者（教員）と対話する形式，学習者同士が対話する形式で引き出すことができるが，学習者同士の対話形式を，2 名の学習者と，3 名以上の学習者間での対話で分け，それぞれ，ペア型・グループ型口頭テスト（paired and group oral tests）と呼ぶことが多い。Ockey and Li (2015, p. 2) によると，口頭コミュニケーションの測定には，以下が主に関わる。

(1) 受験者の口頭コミュニケーション能力
(2) タスク形式
(3) 使用するテクノロジー
(4) 対話者の個人的特徴
(5) 採点者，採点者が使用する採点基準

まず，(1)～(4)が影響して発話（スピーキングパフォーマンス）が生み出される。その発話と(5)が影響した形でスコアが付与される。学習者対話型テストでは，(3)は録音機器等の基本的なテクノロジー以外はあまり関係せず，そのかわりに(4)対話者の個人的特徴が影響しやすいと言われている。

本稿では，学習者対話型テストの研究の流れとその形式の利点と弱点をまとめ，利点の一つである，測る能力（構成概念）の独自性に焦点を当て，学習者対話型テストの可能性を示したい。

2. スピーキングテストの歴史

第二言語のスピーキングテストの実践や研究をたどっていくと，図1のように，(a)面接型から学習者対話型へ(b)直接（direct）形式から半直接（semi-direct）形式へ，(c)同じ場所で受けるテストから，テクノロジーを利用して別な場所で受けるテストへという流れが見えてくる。2.1 で(b)と(c)の概略を述べ，本稿の焦点である(a)について2.2 で詳細に述べる。

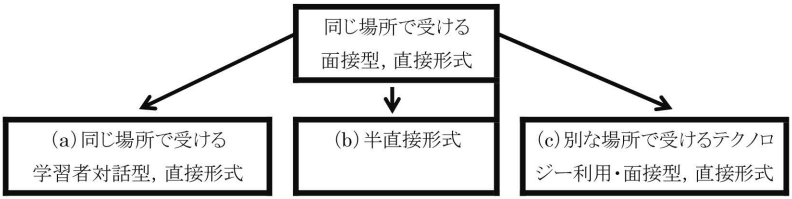

図1 スピーキングテスト形式の変遷

2.1 半直接形式とテクノロジー利用・面接型テストの歴史

スピーキング（口頭）テストの原形は，面接者と対話する直接形式の面接型である（Fulcher, 1997; Sandlund, Sundqvist, & Nyroos, 2016）。面接型は，発話を直接引き出すために妥当性が高くなりやすい一方，対面で実施のために時間もコストもよりかかり，実現可能性（practicality）の点で課題があった。その実現可能性を高めるために開発されたのが，(b)半直接形式のテストである。この形式では，面接型では面接者が口頭で出題するタスクを前もって録音しておき，受験者はそれを聞いて電話やコンピュータに向かって話し，その発話を録音する形である。半直接形式にもバリエーションがあり，面接型がAmerican Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) だとすると，その半直接テストには，Simulated Oral Proficiency Interview (SOPI) というテープに録音する形式や，Oral Proficiency Interview by Computer (OPIc) というコンピュータ上で受験者がアバターと話す形式，Telephone Standard Speaking Test (SST) と呼ばれる電話を使って実施する形式などがある。直接形式と半直接形式を比較した研究によれば，2つの形式でのスコアの平均点は近く，相関も高いことが多いこと，詳細な発話分析では談話の違いがどこかの観点では見られることは分かってきている（Kiddle & Kormos, 2011）。半直接形式で会話を行うことは通常難しいが，最近はコンピュータ実施の半直接形式で，テクノロジー（チャットボット，人工知能，人工会話物，口頭対話システム [chatbots, artificial intelligences, artificial conversation entities, and spoken dialog systems]）を使って疑似的な会話を行うことも試みられている（Ockey, 2017）。

(c)については，面接者と受験者が直接やりとりする形式を保ちつつ，電話やテレビ電話，Skypeなどを用いて異なる場所にいても会話ができる形に修正したものである。この変化は(b)よりは小さいように予想されるが，談話などを分析すると若干の違いが出てくる。例えばNakatsuhara, Inoue, Berry, and Galaczi (2017) では，インターネット経由のビデオ会議システムをタブレット上で使用したInternational English Language Testing System (IELTS) スピーキングテストと，対面型の同じテストを比較した。32名の受験者が両方の形式を受け，その総合的・分析的スコアと使用した発話機能，面接者のテスト実施と採点の行動に違いがあるかを調べた。その結果，スコアには有意な差がなかったが，受験者が使った発話機能と，面接者の実施・採点行動に違いが若干見られた。例えば機能の点では，「明確に話すよう依頼する」機能がビデオ会議型でより多く見られた。さらに，別な場所にいるがウェブを使ったバーチャル環境で学習者対話型テストを行った研究も出てきている(Ockey, Gu, & Keehner, 2017)。テクノロジーの進化に伴い，このようなテスト形式とその実施は今後増え，関連する研究もさらに増加すると思われる。

(a)～(c)ではテスト形式の変遷を扱ったが，使われるテストタスクも多様になってきている。一般的には，スピーキングタスクは現実世界でのタスク(目標言語使用タスク：target language use task)に近く，真正性(authenticity)が高いものが望ましいとされるが，タスクの真正性よりも，口頭コミュニケーションの実際のプロセスで使う，基盤となる能力を測ることを意図して真正性の低いタスクが設定されることもある。その典型がVersant®スピーキングテスト(http://www.versant.jp/versant_test.html)であり，音読や文復唱，口頭語句並び替え，聞いた話の要約などのタスクを使って，リスニング能力も含めて基底能力を測っている(理論的背景については鈴木, 2017; Van Moere, 2012参照)。さらに，スピーキングを単独で測るのでなく，聞いたり読んだりしたものに基づいて話すことを求める技能統合型タスクも広まってきた(佐藤敬典, 2017参照)。タスクの多様性が増えるにつれて，意図した能力をより適切に測るために必要なタスクの特徴や望ましい採点基準などを調べる研究が次々に行われてきている(Cumming, 2014; Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008; Plakans, 2013等参照)。

このようにスピーキングテストでは，受験者から発話を直接引き出す形式は維持しながら，多様な形式やタスクを開発し，その検証を重ねながら，実践と研究の両面で発展してきた。他技能と比べて発展の形が特徴的であるのは，スピーキングテストでは実施や採点の点で実現可能性が低くなりがちで，それを克服すべく様々な試みがなされてきたためであろう。

2.2 学習者対話型テストの歴史

面接型は，現実世界でのタスクと類似しているために真正性が高く，実際に話すことでスピーキング能力を測る形式として，非専門家にも分かりやすい妥当性の高い形式として使用されてきた。しかし，1980年代後半に批判が高まった。面接型では，受験者のスピーキング能力によって提示する質問やタスクを変えるために，タスクと能力を分けて考えられず，妥当性検証を行いにくい点(Bachman & Savignon, 1986)や，面接型での発話が，テストでない普通の会話と異なっていて不自然なものが多い点(van Lier, 1989)等が指摘された。面接型の発話の不自然さについては，

会話で重要とされる「反応性が高く，どの方向にも会話が向く（相互の不確実性）という特徴がなく」（“reactive and mutual contingency”; van Lier, 1989, p. 501），面接者は受験者より強い立場であるために（Johnson & Tyler, 1998）「やりとりが非対称で疑似社会的であり」（asymmetrical and peudosocial; van Lier, 1989, p. 501），面接型の発話は会話と異なるとされた。また，面接者の個人的特徴に発話が大きく影響されうることも実証された（Brown, 2003）。これらの批判により，スピーキングテストの妥当性検証をより厳密に行ったり，テスト中に実際に起こっている発話を詳細に分析したりする流れができ，Staples, Laflair, and Egbert（2017）のように現在でも検証が続いている。

面接型テストの弱点が指摘された後，学習者対話型テストは面接型テストが苦手とする対話能力（interactional competence）をより測れる形式として使用されてきた。学習者対話型テストにおいては，「友達について話しなさい」や「北海道の動物園への旅行に何を持って行くかを4つ決めなさい」などのようにテーマを与えてチャットやディスカッションを求める形式から，ロールプレイ形式で「A: 自宅での夕食に誘いなさい。断られたら都合がよい日を尋ねなさい」「B: 都合が悪いと断りなさい」など話す内容をある程度指定するものまである。

「対話能力」とは，「ある文脈において適切で実際的な使用を行いながら，交互に話し，会話を始め・終え，相手の発話に反応し，トピックを交渉し，発展させる能力」（Ockey & Li, 2015, p. 5）である。Celce-Murcia（2007）が提案した包括的な「コミュニケーション能力」（communicative competence）には，以下の 6 種類の能力があり，対話能力はその中で重要な 1 要素として位置づけられている。

(1) 言語能力（linguistic competence; 文法・語彙・発音等）
(2) 談話能力（discourse competence; 2 文以上を理解・産出）
(3) 方略的能力（strategic competence; 問題を解決，計画・評価）
(4) 社会文化的能力（sociocultural competence; 状況に合わせた表現使用）
(5) 定型表現に関する能力（formulaic competence）
(6) 対話能力（interactional competence; やり取りの中で，言語を使ってどう伝えるか。例：会話を開始・終了する，相手と協力して会話を発展させる，説得する。行動的・会話的・非言語的な［actional, conversational, non-verbal］要素も関わる）

Ockey and Li（2015）によると，学習者対話型テストは，対等な力関係の受験者同士の会話を引き出し，複雑なやりとりに従事する能力を受験者が示す機会となる（p. 10）。このテスト形式では，対話能力を測るためにテスト形式で課されている統制がややゆるく，対等の地位の他人と協力する機会が多くある。しかし同時に，ゆるい統制のために，対話者（他の受験者）の熟達度や性格，親しさなど，多くの要因の影響を受けやすい形式である（Ockey & Li, 2015, p. 10）。この要因と影響の度合いを特定するために，多くの研究が行われてきた（まとめは Sun, 2014 参照）。例えば，対話者の熟達度レベルについて調べた研究が行われ，高い能力の人と組むと，受験者はより多く話し，スコアが高くなるという結果と，スコアに変化はないという結果が報告されている。このように結果が不一致のことも多く，調整変数の統制をより厳密に行ったうえで再検証が必要である（Sun,

2014)。

学習者対話型テストの研究の進行とともに，その形式における発話は，学習者で共に作り上げた会話(co-construction, joint construction)であり，学習者ごとに別のテスト結果を出すことが適切か(McNamara, 1997)について議論が起こり，まだ決着を見ていない。

また，対話能力を含むテストが測る構成概念をどのように捉えるかについて多様な見方が提示される中で，学習者対話型テスト研究からの知見に基づく貢献もなされた。具体的には，構成概念の捉え方は，(a)特性的定義(trait definition)，(b)行動主義的定義(behaviorist definition)，(c)相互作用的定義(interactionalist definition)という 3 つの分類が Chapelle (1998) によって示された。(c)相互作用的定義とは，実用主義的な世界観から，テストでの発話は，学習者の特性と，テスト状況と，両者が関係しあって生まれたものと捉える見方である。その後，Bachman (2007) によって(c)の細分化がなされ，(c1)相互作用を最小限に考慮した定義(Chapelle, 1998)，(c2)相互作用を中程度に考慮した定義(Chalhoub-Deville, 2003)，(c3)相互作用を強く考慮した定義(He & Young 1998)が提案された(詳細は小泉, 2016 参照)。この中で(c3)の構成概念の捉え方は，学習者対話型テストによる発話研究に主に基づき提案されたものである。

現在も，学習者対話型テストにおける構成概念や，ペア型とグループ型における違い等，さらなる精緻化を目指して研究が進んでいる。

3. 学習者対話型テストの利点と弱点

Galaczi and ffrench (2011), O'Sullivan, and Green (2011) によると，主な利点は以下 5 点である。

(1) 学習者対話型テストでは，共に作り出す会話(co-constructed discourse)や，上下関係や助けがあまりない状況での対等な会話が引き出せる

(2) モノローグや面接型では見られない対話能力が見られ，テストで引き出せる発話機能(speech functions)が幅広い(例：意見・根拠を述べる，意味を交渉する，会話を修復する，説得する，トピックを替える; Taylor & Wigglesworth, 2009)

(3) テストで受験者が不安やストレスを感じにくい

(4) 面接型とは異なり，試験官は主に評価に専念できる。2 人以上を同時実施することで時間短縮が可能である

(5) 授業内タスクに似たタスクにできることから，指導との関連を密にすることができ，学習へのプラスの波及効果が得られやすい(Ducasse & Brown, 2009)。学習者間のやり取りを学習活動で行う利点は第二言語習得研究で調べられており(佐藤匡敏, 2017)，以下 4 点である。(a)学んだ知識をアウトプットし，知識を修正する機会が増え，コミュニケーション能力が高まる，(b)コミュニケーションでの問題が起きた時にフィードバックを受け，発話を修正しやすい，(c)最近接発達領域(Zone of Proximal Development)において，足場かけ(scaffolding)を受けて学習が進みやすい，(d)お互いを助け合う関係があると協力的なやり取りが起こりやすく，それにより学習が進みやすい。学習者対話型テストを行うことで，

学習活動としてペア活動が行いやすくなり，その結果，上記の学習へのプラスの影響が期待できる

一方，他形式と比較したときの弱点は以下 6 点である。

(1) 統制を強くかけられる面接者との対話と異なり，学習者間の対話では，影響する要因がより多い（例：熟達度，受験者と対話相手の性格，対話する相手との親しさ，対話者の人数，性別）

(2) 発話が予測できないことが多いため，採点の信頼性を高く保つことがより難しい

(3) 2 人以上で作り上げた会話を評価するため，1 人の能力として解釈しにくい

(4) スピーキング能力に加えて，基礎的なリスニング能力も関わる

(5) 一人あたりの発話時間が面接型より短い（Foot, 1999）

(6) 英語力が低い場合に使いにくい（Foot, 1999）

どの形式においても利点も弱点もあるが，Galaczi and ffrench (2011) や O'Sullivan and Green (2011) は，学習者対話型テストにおける利点は弱点よりも大きいとし，使用を推奨している。学習者対話型テストは，(5)の指導への波及効果を期待するためだけでなく，(1)と(2)のより自然な発話を引き出し，本来の対話能力を測ることのできるテスト形式として注目されている。多様な発話機能を引き出し，対等な会話ができる能力を見たいのであれば，面接者とのロールプレイを行って直接引き出せばよいと考える人もいる。確かにそれは一つの方法であるが，ロールプレイのように，ある機能を使うことが強く求められる状況で発話機能を使える能力を見ていることになる。学習者対話型テストでは，特定の機能使用を強く求められていない状況で，自分から適切な機能を選択して使うことができる能力を見ることができ，その能力こそ，現実でのコミュニケーションで必要な能力と考えられる。実例だが，英会話のグループレッスンで言葉数が少なく，ほとんどやり取りをしない人と組んだことがある。その人は，ロールプレイで交渉役を引き受けたとき，そうでないときの会話と 180 度違う話し方で，質問をし，相手の矛盾点を指摘し，妥協点を探るという難易度が高い発話機能をこなし，交渉役の役割を十分に果たしていた。しかしロールプレイが終わるとまた元の受身の会話に戻ってしまい，不思議だと感じたことがある。この人の場合，面接型では十分な対話能力を示すだろうが，そのテストスコアが現実世界での力を反映していない可能性が高い。その点，学習者対話型テストだと，テストと現実世界のギャップがより少ないと予想される。もちろん，テストだとスイッチが入ったように積極的な対話を行い，現実ではそうでない場合もあるため，テストと現実世界でのギャップというのが存在することは避けられないことである。

4. 学習者対話型テストが測る能力

本形式の主な利点である「測れる構成概念の独自性」(3. の(1)と(2)に対応)について，本節ではさらに詳細に述べる。この観点について知るために様々な研究がなされているが，それは3つのアプローチに分類できる。第1に，テストスコアや発話特徴をモノローグ型や面接者との対話型と比較し，構成概念の相違点を調べた研究である。第2に，学習者対話型テストの中で使われているやり取りのパターンや発話機能を詳細に分類し，特徴を探った研究である。第3に，学習者対話型

テストにおける発話の中で，採点者がどのような点に注目して採点しているかを調べた研究である。これらの研究により，学習者対話型形式の独自性が示されてきている。

4.1 モノローグ型や面接型とテストスコアや発話特徴を比較した研究

第1の例として，Ahmadi and Sadeghi (2016) がある。この研究では，23人のイラン人学習者がモノローグ型・面接型・グループ型テストのすべてを，順序を変えながら受けた。トピックは統制して共通のものを使った。モノローグ型では1人で5分間話し，面接型では7～8分間面接者とやり取りしながら話した。グループ型では3～4名のグループを作り，15～20分間話す形式だった。その結果，スコアの比較では，総合的と分析的スコアともに，グループ型の方が面接型より高い値になった。正確さ（エラーのない節の割合で測定）では，グループ型の方がモノローグ型・面接型よりも高く，より正確な発話をしていた。統語的複雑さ（ASユニット当たりの節数・ASユニット当たりの語数で測定）では，モノローグ型の方が面接型・グループ型よりも高く，より統語的に複雑な発話になっていた。このような違いが出た理由として著者らは，グループ型は，能力をより効率よく示す機会があり，コミュニケーションでのプレッシャーが強かったため，より正確で統語的にシンプルな発話をしたと解釈している。似たような形でスコアや言語的特徴を形式間で比較した研究は，Brooks (2009), Negishi (2015) などがある。Wang (2014) は，過去の研究結果を統合するメタ分析という手法を用い，ペア型とモノローグ型の間にスコアの点で非常に大きな違いがあったことを示した（ペア型またはグループ型 > 面接型; $d = 2.29$, 95%信頼区間: 1.93－2.65, $k = 9$）。

Ockey, Koyama, Setoguchi, and Sun (2014) では，222名の日本人大学生がグループ型・モノローグ型（絵と図の描写）・プレゼンテーション型・モノローグ型（TOEFL iBTスピーキング）テストを受け，そのスコアを比較した。その結果，テスト方法間の相関は中程度（$r = .67$– $.76$）で，形式によって引き出せる力が違うことが示された。モノローグ型（TOEFL iBT）スコアと，他形式の分析的スコアの間の相関は .50～ .74であり，プレゼンテーション型の語彙・文法との相関が最も低く，グループ型の流暢さとの相関が最も高く出たが，全体で中程度の関係ということは一致していた。

このように，スコアでも正確さなどの詳細な観点においても，学習者対話型と他の形式には違いがあり，それぞれに特徴があることは示されてきている。そのため，スピーキング能力の測定においてはより多くの形式を取り入れることが必要であり，限定された形式で測る場合には，それに基づいたスコアの意味が限定されていることに留意する必要がある。

4.2 使用される発話機能・やりとりのパターンを分析した研究

Galaczi (2008) では，Common European Framework of Reference (CEFR) のB2レベル程度のテスト（ケンブリッジ英検のFirst Certificate in English: FCE）におけるペア型でのやり取りのパターンを分類した。FCEテストを受けた60名の受験者の発話は会話分析され，パターンは以下の4つに分類された。

(1) 協力型（collaborative; 高い相互性と対等性が特徴 [high mutuality and equality]）

(2) 非対称型（asymmetric; 一方が独占して話し，一方が受け身で話が限定的）

(3) 混合型(blend; (1), (2), (4) のパターンの特徴が混在)

(4) 平行型(parallel; 相手が始めたトピックを発展させずに，お互いに一人で話す。低い相互性と高い対等性が特徴 [low mutuality and high equality])

この4パターンごとに，テストでの「やり取りを伴うコミュニケーション」(Interactive communication)のスコアを比較したところ，最も高いスコアを得たのが(1)協力型であり，次が(2)非対称型または(3)混合型，最後に(4)平行型となった。筆者の経験では(4)のパターンの会話をする日本人大学生は多いが，そのパターンだとスコアが低くなる傾向があるようである。また Nitta and Nakatsuhara (2014) では，日本人大学生においてペア型テストを長い準備時間ありで行うと，(4)平行型になりやすかったことも報告されている。

その後Galacziは研究を発展させ，Galaczi (2014) では，CEFRのB1～C2間の学習者のペア型テストでの発話を比較した。各レベルで平均的な発話の計41のペア(各レベルで7～12組のペア)のビデオが選ばれ，書き起こしがなされた。その結果，熟達度レベルで大きな違いが3点見られた。

第1に，トピックを発展させる構造が異なっていた。特に，上級者になると，トピックを発展させる度合いが強くなり，自分が始めたトピックでなく相手が始めたトピックを発展させる傾向があった。

第2に，聞き手としてどのようにサポートするかという動きが異なっていた。特に，*yes*や*hm*などの相槌が低い熟達度レベルではほとんどだったのに対し，*absolutely*や*exactly*のような理解したことを示す発言が上級者になるにつれて増えていった。

第3に，話者交替 (turn-taking) のマネジメントに違いが見られた。特に，オーバーラップ・ラッチ (latch) の後の話者交代は熟達度が上がると増加し，逆にギャップ・ポーズがあった後の話者交代は熟達度が上がると減少傾向にあった。全体として，熟達度が上がるにつれて対話能力も向上していくことが示唆された。

学習者対話型は，引き出せる発話機能が多様であることが長所の一つだが，実際に多様な発話機能が観察できるかを調べた研究もある。He and Dai (2006) では，中国のCollege English Test-Spoken English Test (CET-SET) におけるグループ型ディスカッションタスクでの48グループの会話を書き起こした。そこで使われた機能を調べたところ，使用が期待される機能の中で実際に使われたものは限られ，一部の機能のみ出現していることが分かった。例えば，75%以上のペアが使用したのが，「同意(反対)する」「意見・情報を求める」のみで，「会話を発展させる」は17%，「意味を交渉する」は10%のみだった。この研究により，「学習者対話型テストは多様な機能を引き出せる」と前提を置くのではなく，実際に引き出せているかを確認し，多くの機能を引き出すために効果的なタスクやトピックを設定するよう注意すべきであることが分かる。その他，発話機能を詳細に分析した研究の例として，Butler and Zeng (2014), Gan (2010) がある。

Koizumi, In'nami, and Fukazawa (2017) は日本人大学生において，ペア型テストでの発話機能を調べた研究である。ペア型テストでの会話で実際に使われている発話機能は何か，またタスクによって使われる発話機能は異なるかを調べ，系統的にタスクとトピックを設定するのに有用な知見を得ようとした研究である。Koizumi, In'nami, and Fukazawa (2016) で日本人大学生の

初級者から上級者に11個のペア型タスクを実施した発話データから，各タスクについて，20組分のペアの発話サンプルを選んだ。また2名の研究者が，各タスクから引き出すのを意図している発話機能を話し合い，決定した。例えば，「旅行に持って行くものを話し合う」タスクでは，表1の下線を引いた機能が使われると予想した。

表1 発話機能の例（O'Sullivan, Weir, & Saville, 2002に基づく）

種類	含まれる機能
情報に関する機能 (informational functions)	*個人的情報*（過去・*現在*・未来）*を述べる* *意見・好み*・根拠*を述べる*　詳細に述べる　提案する 比較する　不満を言う　推測する　*記述する* 段階を踏んで述べる(staging)　要約する
やりとりに関する機能 (interactional functions)	同意する　反対する　修飾する（modifying） 意見・*情報を求める*　コメントする 許可を求める　あいさつをする　感謝・謝罪する 説得する　会話を修復する 意味を交渉する（理解を確認する，*理解を示す*，共通の土台・目的・方略を確立する，明確に話すよう依頼する，*その依頼に応える*，他人を訂正する・助ける）
やり取りのマネジメントをする機能[a]	会話を始める　トピックを変える　決定する やり取りの中で会話を発展させる（そのために責任を担う，reciprocating）　会話を止める

注：[a](managing interaction functions)。下線は，旅行タスクで使用が予想される機能の例。イタリックは全体で71%以上で使用されていた機能。網掛けは全体で30%未満のみで使用されていた機能

発話を書き起こし，表1のリストを使ってコーディングを行った。表1では簡略化して提示したが，リストには発話機能は35個あり，言語として明示的に表現されているもののみをコーディングした。コーディングの半分は，2名の訓練を受けた者が行い，一致度は89%で十分高かった。その後残り半分を1名がコーディングした。この研究では受験者ごとの分析はせず，タスクごとに使われた機能がどれかを調べた。1つの発言に対して，複数の機能を割り当ててよいとし，使用された割合を計算した。

その結果分かったのは，全体的に発話機能の出現は限定的であることだった。全体で71%以上引き出されていたのは6機能のみだった（その6個はイタリックで表1に提示）。この6機能は，ペア型テストで特にタスクやトピックを工夫せずとも使用頻度が普通高くなるものと考えられる。一方，全体

で30%未満しか引き出されていなかった機能は17機能あり（網掛けで表1に提示），やりとりに関する機能が特に多かった。この17機能は，タスクやトピックを工夫して自然な形で会話に出てくるような仕掛けを考える必要がある。しかし，上で述べたように学習者対話型テストの利点の一つは，自然な形で様々な機能を引き出しうることで，その利点が損なわれないように留意しながら行う必要もある。

事前に設定した，タスクごとにどの程度意図した発話機能が引き出されているかについて調べたところ，多くの場合（76%）意図した機能は引き出されていた。しかし，意図したタスクで出現が不十分だった機能もあり，タスクの中に自然な形で機能が引き出せるように修正する必要があることが示唆された（例: 旅行に持参する物の話し合いタスクで「反対する」「詳細を述べる」，夕食に誘うタスクで「理由を述べる」は意図したが出現割合が小さかった）。

また発話機能とタスクタイプの関連については，ロールプレイではディスカッションよりも「個人的情報（未来）を述べる」「提案する」「感謝・謝罪する」という3機能が多く引き出された。ディスカッションはロールプレイよりも「やり取りの中で会話を発展させる」機能が多く引き出されていた。これらの機能はタスクの指示に入っていなかったが会話で使われており，ロールプレイとディスカッションでそれぞれ引き出しやすい機能だと考えられる。

Koizumi et al. (2017) では，言語として明示的に現れた機能のみをコーディングしたが，無言で反対の意を示すなど間接的に表現されていた機能や，日本語で表現した機能についても今後コーディングする必要がある。さらに同じタスクを母語で行ってどのような会話をするかのデータを得た上で，機能使用の少なさが英語熟達度の低さのために起こるのか，母語でのやり取りも苦手なのかを調べることも求められる（似た観点を分析したButler & Zheng, 2015参照）。2016年の「国語に関する世論調査」から示されたこととして，日本語での対話において意見が違った時に「なるべくことを荒立てないで収めたい」と答える人の割合が増加傾向にあり（1997年では50.7%，2008年では51.3%，2016年で61.7%），「納得がいくまで議論したい」と答えた人（2016年で24.9%）よりもかなり多いことがある（文化庁, 2017）。このため，母語での会話でも表1にあるような「反対する」「他人を訂正する・助ける」「明確に話すよう依頼する」などの「やりとりに関する機能」をあまり使っていない可能性があり，母語と第二言語両面を念頭に置きながら分析を進める必要がある。

4.3 採点者がどのような点に注目して採点するかを調べた研究

テストスコアには，テスト形式によって引き出される発話機能や言語的特徴だけでなく，採点者が発話をどのように捉えたか，どこに注目したかも通常反映される。そのため，テストの構成概念を考える際には，採点者の観点も重要であり，そのような研究も発表されてきている。

Ducasse and Brown (2009) では，17組のスペイン語初級者のペア型会話を採点する時に考えていたことを12名の採点者に語ってもらったところ，以下3点について採点者が注目していることが分かった。

(1) 非言語的対人コミュニケーション（例：視線がどこにあるか・ボディランゲージを伴うか）

(2) やりとりにおけるリスニング（interactive listening; 相手の話を聞いているときに注意して

いるか，聞くことに従事しているか。例えば話を続けるように促すような理解を示しているか，相槌を打っているか）

(3) やり取りのマネジメント（例：話者交代やトピックの一貫性に関するマネジメント）

May (2011) では，12 組のペア型会話を 4 名の採点者が採点し，目立った要素について報告してもらった。その結果，以下 2 点に主に影響されることが分かった。

(1) 互いに協力して到達した度合い（mutual achievement; 相手のメッセージを理解し，反応し，協力的に進め，真正性の高いやり取りと高いやり取りの質に貢献したかなど）

(2) 非言語的なスキル（アイコンタクトを持つ・保つ，強調点をジェスチャーで示すなど）

これらの研究をまとめたSun (2014) によると，研究によって分かった採点者の注目点は，多くが採点基準に記述がないもので，今後採点基準の改良や採点者トレーニングなどに使うことができるか検討すべきである。

4. 終わりに

本稿では，学習者対話型テストが測定する構成概念を中心に，この形式の長所・短所を含めて議論になっている点を扱った。学習者対話型テストについて，本稿がカバーできなかった点もあり，藤田 (2016)，Kley (2015), Sun (2014), Taylor (2011), Van Moere (2013) 等を参照してほしい。

筆者が知る限り，学習者対話型テストでの発話が入った大規模コーパスはない。しかし，このテストを使って第二言語学習者の対話の言語的特徴を記述することは意義深く，今後取り組むべき課題ではないだろうか。ただし，上で述べたように，学習者対話型テストには弱点もある。コーパス化を行うにあたっては，その弱点が強調されないように，影響する要因（調整変数）について可能な範囲で統制・記述する，得られた発話やスコアの解釈時に，調整変数による影響の可能性も含めて解釈する等の注意点を認識しておきたい。

謝辞

本研究は科学研究費，基盤研究(C), No. 26370737 の補助を受けて行ったものである。共同研究者である印南洋先生と深澤真先生と，研究にご協力いただいた中井貴大先生に感謝したい。

引用文献

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, *13*, 341–358. doi:10.1080/15434303.2016.1236797

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). University of Ottawa Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language

proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70,* 380–390.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26,* 341–366. doi:10.1177/0265532209104666

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1–25. doi:10.1191/0265532203lt242oa

文化庁（2017）「平成28年度『国語に関する世論調査』の結果の概要」 Retrieved from http://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/kokugo_yoronchosa/pdf/h28_chosa_kekka.pdf

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessment. *Language Assessment Quarterly, 11,* 45–75. doi:10.1080/15434303.2013.869814

Butler, Y. G., & Zeng, W. (2015). Young learners' interactional development in task-based paired-assessment in their first and foreign languages: A case of English learners in China. *Education 3-13, 43*(3), 292–321. doi:10.1080/03004279.2013.813955

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. Alcón Soler & M. P. Safont Jordà (Eds.), *Intercultural language use and language learning* (pp. 41–57). Heidelberg, Germany: Springer Netherlands.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20,* 369–383. doi:10.1191/0265532203lt264oa

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press.

Cumming, A. (2014). Assessing integrated skills. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. I: Abilities, contexts, and learners; Part 2: Assessing abilities, pp. 216–229). West Sussex, UK: John Wiley & Sons. doi:10.1002/9781118411360.wbcla131

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the *Versant for English* Test: A Response. *Language Assessment Quarterly, 5,* 160–167. doi:10.1080/15434300801934744

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26,* 423–443. doi:10.1177/0265532209104669

Foot, M. (1999). Relaxing in pairs. *ETL Journal, 53,* 36–41. doi:10.1093/elt/53.1.36

藤田智子（2016）「スピーキングの対話の評価」 渡部良典・小泉利恵・飯村英樹・髙波幸代（編）『日本言語テスト学会誌 20 周年記念特別号』(pp. 112–115) 日本言語テスト学会.

doi:10.20622/jltajournal.19.2_0 Retrieved from https://www.jstage.jst.go.jp/article/jltajournal/19/2/19_0/_article

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7)*: Language testing and assessment* (pp. 75–85). Dordrecht, the Netherlands: Kluwer Academic.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5,* 89–119. doi: 10.1080/15434300801934702

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

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

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*, 585–602. doi:10.1177/0265532210364049

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, *23*, 370–401. doi:10.1191/0265532206lt333oa

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Amsterdam, the Netherlands: John Benjamins.

Johnson, M., & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 27–51). Amsterdam, the Netherlands: John Benjamins.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, *8*, 342–360. doi:10.1080/15434303.2011.613503

Kley, K. (2015). *Interactional competence in paired speaking tests: Role of paired task and test-taker speaking ability in co-constructed discourse*. Unpublished Ph.D. dissertation, University of Iowa, U.S. Retrieved from http://ir.uiowa.edu/etd/1663/

小泉利恵(2016)「テストの使用目的と構成概念」渡部良典・小泉利恵・飯村英樹・髙波幸代(編)『日本言語テスト学会誌 20 周年記念特別号』(pp. 50-55) 日本言語テスト学会. doi:10.20622/jltajournal.19.2_0 Retrieved from https://www.jstage.jst.go.jp/article/jltajournal/19/2/19_0/_article

Koizumi, R., In'nami, Y., & Fukazawa, M. (2016). 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. (2017, June). *Examining the construct of paired oral tasks through analysis of elicited speech functions.* Presented at the 4th International Conference of the Asian Association for Language Assessment (AALA), National Taiwan University, Taipei, Taiwan.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*, 127–145. doi:10.1080/15434303.2011.565845

McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, *18*, 446-466. doi:10.1093/applin/18.4.446

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*, 1–18. doi:10.1080/15434303.2016.1263637

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.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing, 31,* 147–175. doi:10.1177/0265532213514401

Ockey, G. J. (2017). Approaches and challenges to assessing oral communication on Japanese entrance exams. *JLTA Journal*, *20*, 3–14. doi:10.20622/jltajournal.20.0_3 Retrieved from https://www.jstage.jst.go.jp/article/jltajournal/20/0/20_3/_article/ -char/en

Ockey, G. J., Gu, L., & Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly*, *14*, 346–359. doi:10.1080/15434303.2017.1400036

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32,* 39–62. doi:10.1177/0265532214538014

Ockey, G., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value*, *7*, 1–21. doi:10.6035/LanguageV.2015.7.2 Retrieved from http://lib.dr.iastate.edu/engl_pubs/74/

O'Sullivan, B., & Green, A. (2011). Test taker characteristics. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 36–64). Cambridge, UK: Cambridge University Press.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19,* 33–56. doi: 10.1191/0265532202lt219oa

Plakans, L. (2013). Assessment of integrated skills. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 205–212). West Sussex, UK: John Wiley & Sons. doi:10.1002/9781405198431.wbeal0046

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass*, *10*, 14–29. doi:10.1111/lnc3.12174

佐藤匡敏（2017）「ペア・グループワークの潜在力を引き出そう」鈴木渉（編）『実践例で学ぶ第二言語習得研究に基づく英語指導』(pp. 89–105). 大修館書店

佐藤敬典（2017）「技能統合型テスト」 小泉利恵・印南洋・深澤真（編）『実例でわかる英語テスト作成ガイド』(pp. 76–79) 大修館書店

Staples, S., Laflair, G. T., & Egbert, J. (2017). Comparing language use in Oral Proficiency Interviews to target domains: Conversational, academic, and professional discourse. *Modern Language Journal*, *101*, 194–213. doi:10.1111/modl.12385

Sun, H. (2014). Paired and group oral assessment. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *14*(2), 68–83. Retrieved from https://tesolal.columbia.edu/article/paired-and-group-oral-assessment/

鈴木祐一（2017）「指導の評価――スキル習得理論の観点から」鈴木渉（編）『実践例で学ぶ第二言語習得研究に基づく英語指導』(pp. 169–183). 大修館書店

Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge, UK: Cambridge University Press.

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

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*, 489–508. doi:10.2307/3586922

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*, 325–344. doi:10.1177/0265532211424478

Van Moere, A. (2013). Paired and group oral assessment. In C. A. Chapelle (Ed.), *The*

encyclopedia of applied linguistics (pp. 4353–4356). West Sussex, UK: John Wiley and Sons. doi:10.1002/9781405198431.wbeal0894

Wang, L. (2014). A meta-analysis of peer-peer interaction in L2 English speaking assessment. *English Teaching & Learning*, *38*(3), 103–137. doi:10.6330/ETL.2014.38.3.04