国際交流基金日本語教育紀要
第19号（2023年）
〔教材開発論文〕

CBT 方式による日本語スピーキングテストの試作と試行

－テストの品質と項目数の検証－

篠原亜紀・夷石寿賀子

〔キーワード〕　CBT 方式、スピーキングテスト、品質、項目数、信頼性

〔要　旨〕

国際交流基金日本語国際センターでは、2019年度より、CBT 方式による日本語スピーキングテストの可能性を探るための調査・研究を進めてきた。2020年度には、CBT 方式のスピーキングテストを 3 セット試作し、その品質と項目数について検証するため、タイの日本語学習者170名を対象に試行試験を行った。その結果、3 つのテストセットにおいて、難易度は0.70～0.74、識別力は0.56～0.57で、セット間に品質の差はほとんどなかった。テストの信頼性（クロンバックのａ係数）は0.834～0.846で、すべてのテストセットにおいて概ね高い信頼性係数が確認でき、信頼性の高いテストであることが立証された。また、試作したテストの項目数が適切であるかどうかを、多変量一般化可能性理論を用いて検証したところ、試作したテストの項目数（インタビュー 4 問、会話 4 問）で高い信頼性（0.823）を得られることがわかった。

1. 背景

外国語教育において、人と人との円滑なコミュニケーションを目標とする学習が重視されるようになり、それに伴い、文法や語彙等の知識を測る従来のテストだけでなく、実際のパフォーマンスを測るスピーキングテストが実施されるようになった。

国際交流基金は、「日本語能力試験」へのスピーキング科目導入の可否を探るため、1998年から2002年にかけて調査と試行試験を実施した（日本語能力試験企画小委員会口頭能力試験調査部会 2003）が、スピーキング科目の増設には至らなかった。また、2014年には「JFS 準拠ロールプレイテスト」（国際交流基金 2014）が公開されたが、プレースメントテスト等の教育現場での活用を想定しているものであり、公式に合否判定やレベル認定をするものではない。

近年、ICT 技術の発展により、CBT（Computer Based Testing）方式のテストが増え、TOEFL iBT、TOEIC Speaking、英検 CBT 等、外国語スピーキングテストで CBT 方式が導入されている。日本語においても、OPIc-J、ONiT、JLCAT といった CBT 方式のスピーキングテストがあるが、英語のような大規模なテストはまだない（篠原ほか 2021）。

CBT 方式のスピーキングテストは、ヘッドセットを装着してコンピューターで受験するも

ので、受験者は音声や画像によって出題された問題に対し、マイクを通して回答する。回答はその場で録音され、後に採点者に送られる。スピーキングテストは、他の技能よりも実施にも採点にも時間と手間がかかる（小泉 2015：55）が、CBT 方式の導入によって、一度に多数の受験者を対象としたスピーキングテストの実施が可能となる。また、CBT 方式は対面式の課題である試験官の差異による影響がなく、公平性を保つことができる。

今後、日本語教育においても大規模スピーキングテストの需要が高まることが見込まれる。国際交流基金日本語国際センターでは、2019年度より、CBT 方式による大規模日本語スピーキングテストの可能性を探るための調査・研究を進めてきた。本稿では、JF 日本語教育スタンダードの考え方に基づいた、課題遂行能力を測る CBT 方式によるスピーキングテストの試作と、2020年度にタイで行った試行について報告し、試作したテストの品質や項目数について検証する。

2. テストの試作

2.1　テストの構成概念

スピーキングテストの試作にあたり、まず、テストの構成概念を策定した。テストのレベルは、海外で学習者層が最も多い等の理由から、Common European Framework of Reference for Languages（以下、CEFR）および JF 日本語教育スタンダード（以下、JF スタンダード）の A2とした。A2のどのような能力を測るかを検討するにあたり、まず、既存の日本語テストや各国語テストの構成概念を洗い出した。その中で、特にスピーキングテストの構成概念を明示しているものや、CEFR 準拠のもの、A2レベル相当のものを中心に情報を収集した。その結果、以下のキーワードが抽出された（表１）。

表１　各種テストの構成概念における A2レベルのキーワード

話題	基本的、個人的、日常的、身近、直接的必要、自分の背景、慣れ親しんでいる、社会に必要な、習慣的な、知っていること （例：家族、買い物、近所、仕事、生活環境、学歴、経歴、友人）
場面	身の回りの状況、日常生活、社会生活、慣れた状況、普通の日常生活
活動	情報交換（やりとり）、説明（産出）、短い社交的なやりとり、短い会話、 簡単に説明、簡単な質問、簡単で直接的な情報交換
ストラテジー	聞き返し、繰り返し要求
質	簡単な文法、単純な文法、センテンス／文で、一般的に使用される表現

さらに、CEFR および JF スタンダードの能力記述も整理し、上記のキーワードとおおよそ一致することを確認した。中でも、CEFR 共通参照レベルの「全体的な尺度」に書かれている A2の記述が構成概念として妥当だと考え、それを引用することとした（表２）。

表２　スピーキングテストの構成概念

A2	・ごく基本的な個人的情報や家族情報、買い物、近所、仕事など、直接的関係がある領域に関する、よく使われる文や表現が理解できる ・簡単で日常的な範囲なら、身近で日常の事柄についての情報交換に応ずることができる ・自分の背景や身の回りの状況や、直接的な必要性のある領域の事柄を簡単な言葉で説明できる （Council of Europe 2004）

なお、「聞き返し」や「繰り返し要求」などのストラテジーについては、A2レベルの言語使用者にとって必要ではあるものの、CBT 方式では測ることが困難だと判断し、構成概念には含んでいない。また、既存の外国語テストでは、読んでから話す、聞いてから話す、といった多技能統合型のタスクを取り入れているものもあるが、本テストでは、スピーキングの能力に限定し、読む能力や聞く能力は測らないこととした。ただし、やりとりをするうえで相手からの簡単な問いかけを理解する能力は含まれる。

2.2　テストの問題構成

構成概念を決定したうえで、それをどのように測るかを検討した。問題構成を決定するにあたっては、TEAP、GTEC、英検 CBT など、既存の CBT 方式のスピーキングテストを参考にした。

表３　スピーキングテストの問題構成と測る能力

パート		項目数	形式・内容	測る能力
1	インタビュー	4	・質問を聞いて、自分自身について答える ・１問につき２つの質問 ・準備時間なし	・個人的情報、直接的関係があることについての情報交換に応じる ・質問に即時に反応し、適切に答える
2	会話	4	・与えられた場面・状況において、イラストで示された内容を話す ・準備時間あり	・仕事や日常生活で直面する場面において、必要なやりとりや、社交のためのやりとりをする ・簡単な情報交換、説明をする

構成は、パート１（インタビュー）、パート２（会話）の２部構成とし、どちらもやりとりの能力を測ることとした（表３）。CBT 方式では、システム上、双方向のやりとりよりも、一人で話し続ける産出タスクのほうが作成しやすいが、ほとんど準備なしでまとまった話をしたり、一人で話し続けたりすることは、A2レベルでは難しいと考えたためである。

項目数は、パート１が４項目、パート２が４項目の計８項目とした。項目数についても既存のスピーキングテストを参考にしているが、CBT 方式に慣れていない受験者の負担や疲労を考慮し、15～20分程度の受験時間が妥当だと考えた。また、項目数が多くなると採点者の人数や採点にかかる経費が増大するため、A2レベルを測るための最低限の項目数にすることを目

指した。この項目数が妥当かどうかについては、検証の結果を4.2.2で記述する。

パート１は、自分自身について答えるインタビュー形式とした。即時応答の能力を測るため、準備時間は設けていない。画面には「面接官」の動画が表示され、その「面接官」からの質問を聞いて答える。質問は１項目につき２つあり、２往復のやりとり（面接官→受験者→面接官→受験者）とした。１つめの質問は、「何」「だれ」「どこ」等の疑問詞を使い、一言でも答えられるものとし、２つめの質問は、「どうして」「どんな」等の疑問詞を使い、回答にある程度の説明が必要なものとした。回答時間は、１つめが８秒、２つめが15秒である。以下は、パート１の一例である（下線は受験者の回答例）。

面接官：好きな動物は何ですか。
受験者：猫です。私は猫が好きです。（８秒）
面接官：それは、どうしてですか。
受験者：猫はかわいいですから。そして、私のうちにいますから。（15秒）

パート２は、場面や状況が書かれた説明文を読んで、イラストで示された内容を話すロールプレイ形式の会話とした。イラストの内容を話すことができ、情報が伝えられたかどうかを測る。準備時間は合計60秒とした。まず、ロールカードのような、状況とタスクが書かれた文が受験者の母語で画面に提示され、読む時間として25秒が与えられる。次の画面ではイラストが提示され、受験者はイラストを見ながら35秒で話す準備をする。その後、会話が始まる。会話はその場面の「相手」の問いかけから始まり、受験者はその「相手」の音声を聞いた後、問いかけに答える形で話す。回答時間は20秒とした。やりとりとして、受験者が話したことに何らかの反応をするのが自然だと考え、1.5往復のやりとり（相手→受験者→相手）になるようにした。以下は、パート２の一例である（斜体は状況を示した説明文、下線は受験者の回答例）。

あなたは今、友だちと明日映画に行く相談をしています。
友だちと待ち合わせの約束をしてください。

友だち：何時に、どこで会いましょうか。
受験者：映画館の隣にカフェがあります。７時にカフェの前はどうですか。（20秒）
友だち：わかりました。

2.3　テストの項目

テスト項目は、JF スタンダードの JF Can-do および JF 生活日本語 Can-do をもとに作成した。作成にあたり、それらの Can-do に基づいて作られた教科書『まるごと 日本のことばと文化』や、『いろどり 生活の日本語』も参考にした。まず、Can-do のリストから、A2レベルのやりとりの Can-do を抽出した。次に、それらの Can-do がどのように実現されるか、上述の教材の中のモデル発話を参考に想定会話を書きだした。その後、想定会話を見ながら、パート 1 またはパート 2 の問題形式に当てはめ、テスト項目を作成した。想定会話から CBT 方式では実現不可能と判断した場合は、テスト項目としての採用を断念した。

項目は、レベルや内容、説明文などについて、作成者以外の者による評価を経て、検討や修正を重ね、テストとして使用可能だと判断したものをアイテムバンクに収めた。アイテムバンクからトピックなどのバランスを考えて組み合わせ、試行のために 3 セットのテストを試作した。

2.4　テストの採点基準

採点は、○△×の 3 段階とし、○はタスク達成で 2 点、△は惜しいがタスク不達成で 1 点、×はタスク不達成で 0 点とした。タスクが達成できたかどうかを中心に評価するが、質的にも A2レベル相当であることが重要だと考え、質的な観点も採点基準に含むこととした。タスクの達成については、まず、○の場合の想定会話をイメージし、発話量や内容、情報量など、どの程度の発話ができればよいかを記述した。質的な観点については、CEFR 共通参照レベルの「話し言葉の質的側面」や、言語能力を示す Can-do を参照し、文法・語彙、正確さ、流暢さ、発音などの観点について記述した。その後、△や×の基準を作成した。課題遂行および質的観点について、分析的に評価するのではなく、総合的に評価して○△×を判定することとした。採点にかかる時間やコストを考慮し、だれでも効率よく、かつ安定して採点ができるよう、できるだけシンプルな採点基準を目指した。

2.5　テストの実装

イラストの作成、動画および音声の収録、説明文の翻訳等を経て、作成したテストを CBT 方式に実装した。テストの実装は業者[(1)]に委託し、画面や動きなどについて業者と相談しながら進めた。テストは、自動遷移とし、一定の時間が経過すると自動的に次の画面に進むようにした。システム上は手動遷移とすることも可能ではあったが、自動遷移は受験者による誤動作を防ぐこともでき、また、ほぼ同じ時間に終了するといった利点がある。各画面には、次の画面に移るまでの残り時間を示すためのタイムバーを設置した。

パート 1 は、受験者が画面上の「面接官」と 2 往復のやりとりをする形式であるが、当初、業者の見解では、そのようなやりとりは CBT 方式では難しいとのことであった。1 ページに

複数の動画／音声を掲載したり、複数の回答を録音したりすることはシステム上不可能であったためである。そこで、1つの音声ファイルに無音区間を入れて2つの質問を入れておき、それを再生し続けることとした（質問①→無音→質問②→無音）。質問音声の再生と同時に回答の録音が開始され、録音は2つめの質問への回答時間が終わるまで続く。つまり、回答データは、質問が再生されている間は無音となるものの、2つの回答が含まれた1ファイルとなる（無音→回答→無音→回答）。図1は、パート1の再生と録音のイメージである。塗りつぶした部分が無音となる。

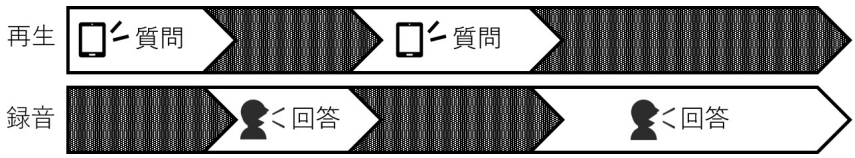

図1　再生ファイルと録音ファイルのイメージ（パート1）

パート2においても、問いかけと反応を無音区間を挟んで録音した1ファイルを再生することで、1.5往復のやりとりを可能とした。

3. テストの試行

3.1　試行の目的

試作したテストの信頼性や妥当性を検証するため、試行を行った。本稿では、テストの信頼性についての検証結果を報告することとする。

まず、テスト項目の難易度や識別力、信頼性係数等からテストの品質を検証する。次に、テストセットの項目数が適切かどうか、つまり、信頼性のある結果を出すために十分な項目数になっているかどうかを検証する。

3.2　試行の概要

3.2.1　日程・場所

テストは、2021年3月19日（金）、20日（土）、21日（日）の3日間、タイのバンコクで実施した。当初、2か国以上での試行を計画していたが、新型コロナウィルスの感染拡大により実施国や実施日程の決定が難航したため、複数国での実施を断念し、実施可能な1か国を検討した。実施国や日程の変更を経て、最終的に、感染状況が比較的落ち着き、多くの日本語学習者からの協力が得られると思われるタイでの実施が決定した。

場所は、カセムバンディット大学ロムクラオキャンパス内にある、CBTの設備が整った業者の試験センターで、各日4セッション（8：00～、10：00～、12：00～、14：00～）に分け

て実施した。

3.2.2　受験者

CEFR／JF スタンダードの A1～B1レベル、または、日本語能力試験（以下、JLPT）の N5未満～N3レベルの日本語学習者を対象として受験者を募った[(2)]。全受験者が 3 セットのテストを受験することとし、申込フォームには、氏名、連絡先、希望するセッションのほか、上述のレベルを記入してもらった。CEFR／JF スタンダードのレベルは自己判断であるが、参考情報として、『みんなの日本語』や『まるごと 日本のことばと文化』の既習課を提示した。

募集の結果、266名の申し込みがあったが（レベル対象外は除外)、当日受験したのは170名（64％）であった。受験者は 3 セットのテスト（セットＡ、セットＢ、セットＣ）をすべて受験することとしたが、全員が同じ順序で受験すると、慣れや疲労等の影響で順序効果が出てしまう可能性がある。例えば、難易度にかかわらず、最初に受けたセットは慣れていないために得点が低い、最後に受けたセットは疲労のために得点が低いといった結果が生じ得る。そのため、順序を入れ替えた 3 つの版を用意し（表 4)、順序効果の相殺を図った。それぞれの版には、自己申告のレベルがバランスよく入るよう割り振った。

表 4　各版におけるセットの順序

版	順序
版 1	セットＡ→セットＢ→セットＣ
版 2	セットＢ→セットＣ→セットＡ
版 3	セットＣ→セットＡ→セットＢ

表 5 は実際に受験した170名のレベル（事前に自己申告）と受験した版である。事前に各版の人数が同じになるよう調整したが、当日の欠席者があったため、版 2 がやや多い結果となった。レベルのバランスは、B1が70名、A1が26名で偏りがあるものの、本テストのレベルである A2の受験者は A2. 1と A2. 2合わせて74名であった。

表 5　受験者のレベルと版

版	A1	A2. 1	A2. 2	B1	計
版 1	7	15	9	22	53
版 2	9	14	19	24	66
版 3	10	7	10	24	51
計	26	36	38	70	170

3.2.3　当日の流れ

表6は、受付から退出までの流れの一例である。ログインから機密保持契約までは手動で進むため、時間には個人差があるが、パート1が始まってからは自動で遷移する。テスト1セットは16分程度である。テスト終了後にはアンケートに回答してもらった。

表6　セッションの流れの一例

時間	流れ
13：45	受付
14：00	入室、ログイン、マイクテスト、注意事項、機密保持契約
14：05〜14：21	セットA
14：21〜14：37	セットB
14：37〜14：53	セットC
14：53〜	アンケート

新型コロナ感染対策として、テスト中もマスク着用を必須とした。事前にマスク着用でも回答が問題なく録音されることを確認した。なお、テストの公平性のため、受験者は持参したマスクではなく、受付で配付された不織布マスクを着用することとした。

3.3　採点

3.3.1　採点者

採点者は、国際交流基金関係者14名に依頼し、協力を得た(3)。採点にあたっては、テスト試行の約1か月前に90分程度の採点者説明会を行い、採点者にテスト内容や採点基準を理解してもらうとともに、ワークショップ形式で採点練習に取り組んでもらった。

さらに、採点者説明会終了後にも採点者に課題を出し、採点練習を行ってもらった。採点者は、まず、採点者説明会で頭の中に作った基準を忘れないよう、説明会直後に3セット分の採点練習を行う。次に、採点開始の直前に、基準を思い出すために1セット分の採点練習を行った。各採点練習において、正しく採点できなかった項目が多かった採点者は、フィードバックを得るとともに、再度、他のセットを採点することとした。また、採点者には、採点開始後や採点の途中で間が空いてしまったときなどにいつでも閲覧／練習できるようなフォームを提供した。

上述のようにして、可能な限り、だれが採点しても同じ結果になるよう（採点者間信頼性）、また、いつ採点しても同じ結果になるよう（採点者内信頼性）務めた。

3.3.2　採点方法

テスト全日程終了後、2021年 3 月22日より採点を開始した。採点期間は 2 週間とし、 1 名の採点者が採点する回答数は受験者36～37名分とした。実際の受験者数は170名であるが、全受験者が 3 セットを受験しているため、実質510名分の回答データがあることになる。それを14名の採点者で分担して採点したため、各採点者はパート 1 を36～37名分、パート 2 を36～37名分採点することとなった。

採点は、業者が所有するオンライン採点システムを使用して行った。受験者の回答データは、テスト終了後まもなく採点システムのデータベースに移行され、採点が可能となる。各採点者は、与えられた ID とパスワードで採点システムにログインし、採点を行う。採点システム内では、回答データだけでなく、指示文、質問の動画／音声、イラスト等の情報もすべて表示される。採点者は回答音声を再生して採点を行うが、質問音声と回答音声を同時に再生することによって、一連のやりとりを聞くことが可能となる。

採点システムのアクセス権限には、「採点者」と「管理者」があり、管理者 ID でアクセスすると、各採点者の採点状況を確認したり、採点結果を修正したりすることができる。採点システムでは、採点者が一度採点すると回答は採点済みとなり、他の採点者に振り分けられることはない。つまり、複数名で同じ回答を採点することはシステム上不可能であり、採点の信頼性が懸念された。管理者 ID を持つ者は採点結果の修正が可能であることから、採点の信頼性を高めるために以下のような方法を取った。

①採点者は採点の際、判断に迷ったり、採点結果に不安が残ったりする場合は、コメント欄にコメントを残す（一次採点）
②二次採点者が管理者 ID で採点システムに入り、コメントのついている回答を聞いて、採点者の採点結果に同意できればそのまま、そうでなければ、採点結果を修正する（二次採点）

二次採点は、テストの試作に携わった筆者らが分担して各自行った。 1 名で判断できない場合は、議論して最終判定を出した。

4.　結果と分析

試行の結果と、テストの品質、テストの項目数について分析した結果を以下に記述する。分析は、業者に委託して行った。なお、受験者170名中、 1 名がマイクの不具合により録音されていない部分があったため、169名を分析対象とした。

4.1　結果

表７は、各セットにおける受験者の得点の記述統計である。テストは１セット16点満点（各項目２点×８項目）で、平均点は11.14～11.82、標準偏差は3.68～3.77であった。最低点は０～１点、最高点は16点で、ばらつきが大きいことがわかるが、これは、受験者をA1～B1レベルの幅で募集したためである。

表７　各セットの得点の記述統計（N ＝169）

セット	平均点	標準偏差	最低点	最高点
セットA	11.63	3.68	0	16
セットB	11.14	3.73	0	16
セットC	11.82	3.77	1	16

4.2　分析

テストの品質を検証するため、各項目の難易度や識別力を算出し、各セットの平均難易度や平均識別力を確認した。また、テストの信頼性の検証として、各セットのクロンバックのａ係数を確認した。

各セットの統計値を出すにあたっては、まず、受験者によって受験順序が異なることが原因で生じる順序効果の影響がないかを確認した。今回順序効果が見つかったのは、セットＡにおける版１と版３の間の識別力のみであり、それぞれのセットに統計値が明確に異なる証拠はなかった。今回の試験デザインでは、各版のデータを統合することにより順序効果を相殺することが期待できるため、データを統合し、各セットの統計値を確認する。

4.2.1　テストの品質

表８～10は各セットにおける項目難易度と識別力である。項目難易度は、それぞれの項目の平均スコアを最大スコアで割ることにより計算される平均スコア取得率で評価している。識別力は、各受験者の総得点とそれぞれの項目の獲得得点の間の相関をとったピアソンの積率相関係数によって評価している。ここではHenrysson（1963）による手法を使用し、相関係数を計算する際に対象となる項目の得点を合計点から除外している。難易度は、０から１の間の値をとり、１に近いほどその項目は易しいと判断できる。各項目を見ると、最小値が0.43で、やや難しい項目はあるものの、各セットの平均は0.70～0.74であった。セットＢがやや難しく（0.70）、セットＣがやや易しい（0.74）が、その差は小さく、標準偏差からも差はほとんどないことがわかる。

識別力（ピアソンの積率相関係数）は、-１～１の値をとり、数値が高い項目ほど能力の高

い受験者とそうでない受験者をはっきりと区別できるといえる。一般的に0.2以上が望ましいと考えられているが、各項目を見ると、最小値が0.47で、セット間の差はほとんどなく、平均は0.56～0.57であった。

表８　セットＡの難易度と識別力

セット	項目	難易度	識別力
セットA	パート1_1	0.69	0.62
	パート1_2	0.58	0.48
	パート1_3	0.55	0.63
	パート1_4	0.81	0.60
	パート2_1	0.77	0.55
	パート2_2	0.70	0.68
	パート2_3	0.88	0.60
	パート2_4	0.85	0.55
	平均	0.73	0.56
	標準偏差	0.12	0.06

表９　セットＢの難易度と識別力

セット	項目	難易度	識別力
セットB	パート1_1	0.74	0.65
	パート1_2	0.55	0.67
	パート1_3	0.78	0.62
	パート1_4	0.68	0.65
	パート2_1	0.43	0.57
	パート2_2	0.81	0.53
	パート2_3	0.82	0.53
	パート2_4	0.79	0.56
	平均	0.70	0.57
	標準偏差	0.13	0.04

表10　セットＣの難易度と識別力

セット	項目	難易度	識別力
セットC	パート1_1	0.74	0.51
	パート1_2	0.66	0.59
	パート1_3	0.63	0.68
	パート1_4	0.73	0.67
	パート2_1	0.68	0.66
	パート2_2	0.79	0.47
	パート2_3	0.87	0.53
	パート2_4	0.84	0.72
	平均	0.74	0.57
	標準偏差	0.08	0.08

表11は、各セットのクロンバックのα係数である。クロンバックのα係数は、テストの信頼性を示す係数で、テストが一貫して同じ概念や対象を測定しているかどうか（内的整合性）を評価する。クロンバックのα係数は－１～１の値をとり、数値が高いほど信頼性が高いといえ、一般的に0.80以上が望ましいと考えられている。各セットのα係数は0.834～0.846で、テストの信頼性は、すべてのセットで概ね高いといえる。

表11　各セットのクロンバックのａ係数

セットA	セットB	セットC
0.834	0.837	0.846

4.2.2　テストの項目数

テストの項目数が測定精度に与える影響の検証を、多変量一般化可能性理論を用いて実施した。一般化可能性理論は、テストで１項目につき何名の評価者を用意したらよいか、受験者１名に何項目与えたらよいか等、効果的なテスト計画を立てるのに必要な情報を与えてくれる（池田 1994）。特にスピーキングのようなパフォーマンステストでは、項目の違いや評価者の違いなどの要因で測定値に誤差が生じやすい。一般的に項目や採点者の数を増やせば誤差が少なくなり、測定は安定するが、実施や採点に時間のかかるスピーキングテストにおいては、測定が安定する最小限の項目数および採点者数を検討する必要がある。一般化可能性理論では、いくつの項目でどの程度の信頼性が得られるかをシミュレーションすることができ、坂野（2008）、森重・渡部（2017）、Koizumi et al.（2019）等でも、スピーキングテストの項目数を一般化可能性理論を用いて検証している。

一般化可能性理論は、一般化可能性研究（Generalizability study：以下、G研究）と決定研究（Decision study：以下、D研究）の２つの段階に分けられる。G研究では、分散成分（誤差の要因となる成分とそのばらつき）が推定され、D研究では、具体的な条件（項目数や評価者数）における一般化可能性が検討される。今回は、試行した採点デザイン（採点者１名＋二次採点者）において、どの程度の項目数であれば十分な一般化可能性係数が得られるかのシミュレーションを行った。

表12は、G研究において推定された分散成分とその割合である。「受験者」の数値は受験者の得点がどのくらいばらついているかを、「項目」の数値は項目の難しさの違いによって生じるばらつきを示す。「残差」は受験者と項目の交互作用とそれ以外の要因による誤差を示している。受験者の分散成分は、パート１が42.14％、パート２が34.32％で、パート１のほうがばらつきが大きいことがわかる。一方、項目は、パート１が5.48％、パート２が12.44％で、パート２のほうが項目のばらつきが大きい。受験者と項目を比較すると、パート１、パート２のいず

れにおいても受験者のばらつきのほうが大きい。受験者の分散成分の割合が高いときは、テストが受験者の能力を測っている割合が高い（平井 2018）ため、好ましい結果といえる。

表12　G 研究における推定された分散成分とその割合

要因	パート 1 （インタビュー）	パート 2 （会話）
受験者	0. 224（42. 14%）	0. 166（34. 32%）
項目	0. 029（ 5. 48%）	0. 060（12. 44%）
残差	0. 279（52. 38%）	0. 258（53. 25%）

表13は、D 研究で算出された一般化可能性係数（G 係数）と信頼度指数（ϕ 係数）の一覧である。G 係数は集団基準準拠テスト（相対評価）に、ϕ 係数は到達基準準拠テスト（絶対評価）に使用するのが一般的である（平井 2018)。本スピーキングテストはA 2レベルに達しているかどうかを測る到達基準準拠テストであるため、ϕ 係数を確認する。表13より、現在のセット 1 つ分（インタビュー 4 項目、会話 4 項目）では0. 823であることがわかる。数値の基準は、テストの性質や社会的影響の程度によって変わるものの、一般的には0. 8以上であれば信頼性が高いといえる（平井 2018)。 3 項目ずつの場合は0. 777であることから、項目を減らすことはできないが、現在の 4 項目ずつ、またはそれ以上で、高い信頼性を得られることがわかる。

表13　D 研究における G 係数と ϕ 係数

パート	各パートの項目数										
インタビュー	12	11	10	9	8	7	6	5	4	3	2
会話	12	11	10	9	8	7	6	5	4	3	2
G 係数	0. 942	0. 937	0. 931	0. 924	0. 916	0. 905	0. 891	0. 872	0. 844	0. 803	0. 731
ϕ 係数	0. 933	0. 928	0. 921	0. 913	0. 903	0. 891	0. 875	0. 853	0. 823	0. 777	0. 699

5. まとめと今後の課題

試作した CBT 方式のスピーキングテストを、海外の日本語学習者を対象に試行することによりその品質を確認した。3 つのテストセットにおいて、難易度は0. 70～0. 74、識別力は0. 56～0. 57で、セット間に品質の差はほとんどなかった。テストの信頼性（クロンバックの α 係数）は、0. 834～0. 846で、セット間に若干の違いはあったものの、すべてのセットで概ね高い信頼性係数が確認でき、信頼性の高いテストであることが立証された。

試作したテストは、パート 1 （インタビュー） 4 項目、パート 2 （会話） 4 項目の計 8 項目からなるが、その項目数が適切かどうかを、多変量一般化可能性理論を用いて検証した。その

結果、現在の項目数で高い信頼性（0.823）を得られることがわかった。各パート5問以上あればさらに高い信頼性（0.85～）が得られるが、受験時間や採点にかかる時間・コストを考慮して項目数を決定する必要がある。

今回の試行の対象はタイの学習者のみで、限られたデータによる検証であった。CBT方式による日本語スピーキングテストを実現させるためには、他の国での試行や、採点者数の検討、合格基準点の検討など、さらなる分析が課題である。

謝辞：スピーキングテストの試作・試行および分析に関してアドバイスをいただいた清泉女子大学の小泉利恵教授、テスト問題作成にご協力いただいた日本語国際センター専任講師の方々、受験者および採点者の皆様、ご協力くださったすべての方々に厚く感謝申し上げます。

〔注〕

(1) テストの実装はプロメトリック株式会社に委託した。採点システムおよび分析も同社に委託している。
(2) 受験者の募集は、タイ国元日本留学生協会（OJSAT）に委託した。OJSATのWebサイトやFacebookページで募集し、申し込みはGoogleフォームで行った。
(3) 採点は、日本語国際センター専任講師、元専任講師、職員、および、国内待機中の国際交流基金海外派遣日本語専門家に依頼した。

〔参考文献〕

池田央（1994）『現代テスト理論』、朝倉書店
小泉利恵（2015）「スピーキングの評価－スピーキングテスト作成・実施を中心に－」望月昭彦・深澤真・印南洋・小泉利恵（編著）『英語4技能評価の理論と実践－CAN-DO・観点別評価から技能統合的活動の評価まで－』、43-57、大修館書店
国際交流基金（2014）『JF日本語教育スタンダード準拠 ロールプレイテスト テスター用マニュアル』、国際交流基金
篠原亜紀・夷石寿賀子・石田華奈子・李文鑫（2021）「CBT方式によるスピーキングテストの現状」『国際交流基金日本語教育紀要』17、227-238
日本語能力試験企画小委員会口頭能力試験調査部会編（2003）『日本語能力試験企画小委員会口頭能力試験調査部会報告：口頭能力試験科目の創設に向けて』、国際交流基金関西国際センター
坂野永理（2008）「一般化可能性理論による日本語口頭プレースメントテストの検討」『日本テスト学会誌』4 (1)、23-32
平井明代（2018）『教育・心理・言語系研究のためのデータ分析－研究の幅を広げる統計手法－』、東京図書
森重里保・渡部倫子（2017）「一般化可能性理論を用いた外国人児童生徒のためのJSL対話型アセスメントDLA〈話す〉の検討」『広島大学日本語教育研究』27、31-34
Council of Europe（2004）『外国語の学習、教授、評価のためのヨーロッパ共通参照枠』、吉島茂・大橋理枝（訳、編）、朝日出版社
Henrysson, S. (1963). Correction for item-total correlations in item analysis. *Psychometrika, 28: 2*, pp. 211-218.
Koizumi, R., Kaneko E., Setoguchi, E., In'nami, Y. & Naganuma, N. (2019). Examination of CEFR-J spoken interaction tasks using many-facet Rasch measurement and generalizability theory. *Language Testing and Assessment, 8: 2*, pp. 1-33.