プレースメントのための日本語スピーキングテスト
――タスクと判定ツールの検証――

ボイクマン総子・根本愛子・松下達彦

要　旨

プレースメント用のスピーキングテストは，大勢が短時間で受験でき，判定が簡便で，信頼性や妥当性の高いテストが理想である。そこで，筆者らはテストタスクと，ルーブリックと音声サンプルによる判定ツールの開発に着手した。テストの信頼性や妥当性を検証するため，開発した判定ツールを用いて初級から上級の受験者32名の「断り」のタスク結果を日本語教師4名に判定してもらう実験を行った。判定結果は，プレースメント時の読解などの受容能力より産出能力を示す作文と，リスニング要素を含むSPOTとの相関が高かったことから基準関連妥当性を一定程度満たしていると言える。判定者間の一貫性や相関も高く，受験者1名あたりの判定時間が1～2分であったことからも，本タスクと判定ツールは，簡便で一定の信頼性も確保できていると言える。ただし，中級の判定は初級や上級より難しいことがうかがわれた。

【キーワード】 スピーキングテスト，プレースメント，タスク，判定ツール，日本語教師

1. はじめに

プレースメントテスト（PT）におけるスピーキング（SP）テストは，一度に大勢が受験でき，実施が簡便で短時間で判定ができ，判定者間で一貫性の高い結果が得られることが望ましい。しかし，PTでSPテストを実施している機関は，文法や漢字テストを実施している機関よりも経験上少ない。SPテストは，機材調達や対面で実施する際の人的リソースなど実用性の面でのハードルが高いためだろう。加えて，現存の日本語SPテストは，信頼性の検証が不十分で判定者に高度なスキルが要求されることも実施困難な要因だと考えられる。

そこで，筆者らは，上述の課題に応える，PTのためのSPテストの開発に着手した。本研究は，開発途上のSPテストのうち，状況対応タイプの「断り」のタスクについて，当該タスクと判定ツールの「有用性」（Bachman & Palmer 1996）を実験により検証するものである。「有用性」とは，「信頼性，構成概念妥当性，真正性，相互性，インパクト，実用性」の6つからなる。信頼性は測定の一貫性，構成概念妥当性とは測りたい能力を測っているか，真正性は目標言語使用領域とテストタスクの特徴が合致しているか，相互性は受験者の個人的特徴がタスクにどう関連するか，インパクトは社会や教育，個人に及ぼす影響，実用性は実施に必要なリソースと使えるリソースの関係のことである。

『日本語教育』175号（2020.4）　ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

2. 先行研究と研究課題

2-1 日本語の SP テストとその研究

日本語 SP テストには，対面式テストの OPI（牧野他 2001）や JF 日本語教育スタンダード準拠ロールプレイテスト（JF-RP）（国際交流基金 2017），電話による Japanese Standard Speaking Test（JSST），コンピュータによる自動判定の SJ-CAT（石塚他 2017）などがある。

各々独自の特長があるが，PT で実施するには実用性と信頼性の面で課題もある。OPI と JF-RP は対面式であるため，一度に大勢が受験できず時間がかかり実用性の面で難がある。また，テスト実施者のトレーニングも必要である。日本語の OPI では判定結果に関する統計的検証が行われた研究は管見の限りないが, JF-RP の統計的検証は, 波多野（2018）が行っている。波多野は，JF-RP のうち B1 と B2 のタスクを課された受験者 45 名のタスクの達成度についてテスト開発者 4 名が 4 段階で判定した結果を検証した。結果，4 名の判定には高い一貫性が見られ，B1 と B2 のレベル間でタスクの難易度に逆転が見られず, テストの最終判定は学習者能力値と高い相関が見られたという。しかしながら，波多野の研究は JF-RP の B1 と B2 のタスクのみを検証したものであることと，PT では実用性の面からできるだけ少ないタスクで初級から上級という広範囲のレベルに受験者を配置する必要があるという点で，JF-RP が PT の SP としては必ずしも適したテストではないと考えられる。

JSST は複数で判定した結果を，リードレイターが確認することで評価の品質が一定に保たれているとされているが，受験料がかかり実用性が低い。SJ-CAT は，日本語教員による採点結果との相関はあるが，テストタスクごとにばらつきが見られたという（石塚他 2017）。さらに，石塚らの検証の結果，JSST と SJ-CAT の各タスクの判定の相関はあまり高くなかったという。なお，テスト実施に 15 〜 30 分かかり，特に，能力の高い学習者に対するテストの所要時間は長くなることから，実用性の面でも課題が残る。

2-2 開発したテストと研究課題

筆者らは，出題タスクを予め録音しておき受験者はそれを聞いて話すという反応型のテストを開発し, STAR（Speaking Test of Active Reaction）と名付け, テストタスクと判定ツールを開発した。本テストの目指す測定領域の構成概念は, Celce-Murcia（2007）に基づき,「『社会文化的能力』『談話能力』『言語能力』『定型表現に関する能力』『対話能力』『方略的能力』を含む包括的口頭コミュニケーション能力」と定義する(1)。現在，STAR では，①状況対応タスク 3 種（断り，勧誘，依頼）と②絵を見てストーリーを描写するタスクを組み合わせている。会話と独話の両方で構成概念妥当性を高める目的のほか，①のみによる天井効果と②による床効果を回避し弁別性を高めるためである。

本研究では, ①の「断り」とその判定ツールである「ルーブリック」と「音声サンプル」について，相互性とインパクト以外の有用性―信頼性，構成概念妥当性，真正性，実用性―を検証する(2)。なお，「断り」のタスクには，Celce-Murcia（2007）の定義する包括的口頭コミュニケーション能力が必要であるため一定の代表性があると考える。

本実験の検証に際しては，教師歴の長い（16 年以上）日本語教師 4 名が判定を行った。本実験の研究課題は，「状況対応タスクの『断り』とその判定ツールは，（1）信頼性があ

『日本語教育』175号（2020.4）
ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

るか，（2）妥当性があるか，（3）真正性および実用性があるか」である。

3. 方法

3-1　実験に使用した SP テストの概要

実験に使用した SP テストの受験者は 32 名で, 2017 年 10 月に状況対応の「断り」と「勧誘」のタスクと，話を聞いてストーリーを再現するタスクを受験した。男性 17 名，女性 14 名，未記入 1 名で，第 1 言語は英語 11 名，中国語 8 名，韓国・朝鮮語 7 名，スペイン語 3 名，ドイツ語，フランス語，タガログ語各 1 名だった。年齢は不明だが，都内国立 X 大学の交換留学生で送り出し校の 3 〜 4 年生だった。レベルは初級から上級まで多様であった。

実施手順は次の通りである。まず，受験者に次の（1）に示す状況をコンピュータ上の画面で提示し，「書かれた状況を読み，自分ならどう言うか想像してみて下さい。相手の発話を聞き終わったら，各自録音を始めて相手の発話に対して返事をしてください。発話は各状況で自分が必要だと考え得る十分な発話を 1 分以内で行ってください」と英語で教示した。受験者は英語能力が高いことが条件である交換プログラムの留学生であるため，教示には英語を用いた。発話を 1 分以内としたのは，SJ-CAT では 45 秒〜 1 分であったこと, 以前の PT 実施経験から 1 分あればタスクが完遂できることが明らかだったためである。

受験者が画面上のスタートボタンを押すと，相手の発話となる音声キュー（2）が流れ，聞き終わると同時に録音がスタートする。受験者はヘッドホンとマイクを使い発話を行うという手順でテストは行われた。所要時間は, 説明を理解するのに 3 分,「断り」と「勧誘」のタスク受験は各 3 分強で，SP テスト全体では 10 分程度だった。なお，X 大学ではコンピュータを使用して実施しているが，携帯電話や IC レコーダーでの録音でも実施可能である。

（1）受験者に提示した状況

> You are having lunch alone at the university cafeteria half an hour before your Japanese class.
> While eating, another student from the music club which you belong to, came to join you at your table.
> You are a third-year student and she is a fourth-year student of the same university.
> You and she have gone out together occasionally, and are good friends.
> Over lunch you began to discuss different types of music and realized that you both have similar taste. After about 20 minutes, you finished eating.
> She suggests skipping the Japanese class and going to a club down the street to hear a new band, but you don't want to go.

（2）音声キュー

> ねえ，下北にあるクラブでさっき話してたバンドがこれから演奏するんだけど，授業

『日本語教育』175号（2020.4）
ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

行くのやめて，一緒に聞きに行かない？

3-2　状況対応タスクの判定基準

ルーブリック（表 1）は，OPI と CEFR（Council of Europe 2018）を参考に，筆者らのうち 2 名が開発した。「課題達成，テキストの型，文法・表現の正確さ」の項目は OPI を，各表現は OPI と CEFR の記述を参考にした。レベル 1 は初級前半，2 は初級後半，3 は中級前期，4 は中級中期，5 は中級後期，6 は上級に相当する。もう一つのツールの音声サンプルの文字化資料は表 2 に示した。なお，判定者には文字化資料は示さず，音声のみを提供した。

3-3　実験手順

本研究の実験の手順は以下の ⅰ）～ⅲ）の通りである。

ⅰ）STAR 開発者の A と B が 32 名の「断り」タスクの結果をレベル 1 から 6 に判定する。

ⅱ）AB とは異なる機関の日本語教師 CD が判定ツールを用いてレベル 1 から 6 に判定する。

表 1　状況対応タスク判定のためのルーブリック

	レベル 1	レベル 2	レベル 3	レベル 4	レベル 5	レベル 6
課題達成	できるが直接的		簡単に理由や状況を述べたうえで直接的な表現で課題を達成できる	理由や状況を述べたうえで直接的な表現で課題を達成できる	理由や状況を述べたうえで間接的な表現も交えて課題が達成できる	
対人配慮	配慮ができない，または，「すみません」や「ちょっと」を使った配慮ができる		課題達成のための最低限の配慮ができる	課題達成のためにそれなりの配慮ができる	課題達成のための十分な配慮ができる	
テキストの型	単語レベルだが，いくつか単文が出てくる	文レベルだが，単文が多い	複文を使って話ができる。一部で段落が見られる	ほぼ段落レベルで話すことができる	段落レベルで話すことができる	
文法・表現の正確さ	基本は単語，文は定型的なもののみ	定型的な文，表現のみ	間違いがあり，聞き手の理解を妨げることがある	間違いがあるものの，聞き手の理解を妨げることはほとんどない	間違いはあるが，あまり目立たない	ネイティブレベルでほとんど間違いはない
文法・表現の豊かさ	決まったもののみ	日常的に必要な最低限の表現が使える	日常的な表現が十分にできる	さまざまな表現を使おうとする努力がみられる	さまざまな表現を問題なく使うことができる	慣用句や比喩なども交えて話すことができる
流暢さと言い直し	定型的な文が長い休止を伴ったり，言い直ししたりしながら言える	言い直しが多い	流暢さがみられるが発音の悪さでわからない部分がある。言い直しも多い	流暢だが，ときどき言い直しがみられる。発音は気になるが，わからないことはない	言い直しを多少交えながら，流暢に話すことができる。少々発音が気になる	流暢で，言い直しがあっても気にならず，発音もよい

『日本語教育』175号（2020.4）
ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

表 2　STAR 実施時の異なるレベルの音声サンプルの文字化資料（発話の通りに記述）

レベル 1	ぜんぜん，大丈夫です。わたし日本語をクラスは，あー，きー，（沈黙）あー，（沈黙）でも私は日本語のクラスではありま，あのーあー，ちょっと少しんーてつだまずいから。
レベル 2	ん，ありがとうございます。でもあー日本語の，クラスが，あー，さんじゅうーに，あ，さんじゅうにあります。ん。一番クラスです。が，わたしは行きます。あー，あー，またね。
レベル 3	あすみません今はちょっとあー，わたしの日本語が下手ですから，んー，日本語のクラス，行かなければなりません。
レベル 4	へーあのわたしの好きなバンドもうコンサートがありますな，ありますか。えー本当に行きたいんけど，そのコンサートはだだその時間だけありますか。んーでもわたし今から 10 分以内授業にありますから，なんか考えちゃうね。日本語のクラスだから行きたいんだ，行きたいと思っていますだから。先生の授業は本当に面白いからなんかさぼって，ほしくないね。次またありますから，せひわたしに教えてくださいね。それじゃまたね。
レベル 5	えっまじで。もう今始めるの。あーでも日本語の授業あるのね。あと 10 分後で。授業は 2 時間だけなので，そのバンドはそのあとまだいるのかな。ま，授業の後すぐに行くので今はこの授業をさぼるのはだめですよ。私は留学生なのでこの授業をさぼったら単位が取れなくて卒業できないんですよ。ですからこのバンドの音楽は大好き，でも卒業できないのはちょっと厳しいなので。うん授業後必ず行きますから。えっとこのバンド終わったら他のクラブでも行くので大丈夫ですよ。それに先生に授業が来なくて叱られたら私も単位もちょっと取れなくなっちゃったのでそれもまずいですね。授業んあとすぐ行くので大丈夫です。すみません。
レベル 6	まそれはちょっと無理かな。まなんかじゅんちゃんさえも無理かなと思って。なんかあのー，アテンダンスレートっていうものがあって，外国人学生にとってめっちゃ大事なルールなんで，まちょっとそうっすねー，でも来週の渋谷のクラブでなんかじゅんちゃんの演出があるので，ま一緒に見に行かない？ん，はい，じゃ私クラスねー，はーい。

iii）A ～ D の 4 名の判定結果につき，判定の信頼性，妥当性，実用性を検証する[3]。

3-4　判定者情報

判定者の属性と所要時間は表 3 の通りで，判定前作業とはルーブリックと音声サンプルの確認作業である。A と B にこの時間がないのは判定ツール開発者であるためで，所要時間の差は聞き直しの有無や回数による。しかし，慣れにより時間短縮は可能だと思われる。

表 3　本実験の判定者と判定の所要時間

	教師歴	OPI 資格	判定前作業	32 人分の判定所要時間
判定者 A	24 年	元テスター	N/A	約 20 分
判定者 B	19 年	現テスター	N/A	約 20 分
判定者 C	22 年	非テスター	約 30 分	約 60 分
判定者 D	16 年	非テスター	約 30 分	約 120 分

4. 結果

4-1　4 名の判定者の判定

絶対一致に基づく各判定者間の Fleiss の Kappa 係数は表 4 の通りで，全体としては中程度の一致と言えるが，個別に見ると比較的よく一致しているペアと相対的に一致度の低いペアがあった。しかしながら，級内相関係数の判定者内信頼性 ICC（1，1）は .867，判定者間信頼性のうち絶対一致に基づく ICC (2, 1) は .868, 相対一致に基づく ICC (3, 1) は .888 で[4]，クロンバック α も .975 と非常に高く，総じて高い信頼性が確保されていた。

『日本語教育』175号（2020.4）
ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

表 4　断りタスク（1 ～ 6 点）の平均・標準偏差と判定者間の Kappa 係数（Fleiss）（n=32）

	平均	標準偏差	判定者 A	判定者 B	判定者 C	判定者 D
判定者 A	3.22	1.56	1	.582	.426	.342
判定者 B	3.47	1.52	.582	1	.320	.265
判定者 C	3.22	1.73	.426	.320	1	.427
判定者 D	2.84	1.35	.342	.265	.427	1
全体	3.19	1.54				

4-2　断りタスクと他の PT との関係

X 大学の PT の実施順序と所要時間は，SPOT バージョン A（小林他 1996）（10 分），漢字（16 分），語彙（25 分），文法（20 分），SP（10 分），読解（30 分），作文（30 分）で，総合判定は，これら全てのテスト項目を考慮して総合的に判定した。「断り」タスクは SP テストの一部で，「断り」タスクの判定結果と他のテスト項目との相関は表 5 の通りである。

表 5　判定者 4 名の判定結果とその他のテストの相関（ρ）（n=32）

	総合判定	SPOT	漢字	語彙	文法	読解	作文
判定者 A	.686***	.795***	.411*	.526**	.777***	.442*	.755***
判定者 B	.664***	.724***	.431*	.541**	.707***	.467**	.682***
判定者 C	.600***	.751***	.367*	.431*	.748***	.312	.712***
判定者 D	.686***	.740***	.474**	.555**	.729***	.482**	.673***

*p<.05, **p<.01, ***p<.001

ここから，漢字や語彙との相関は相対的に低い一方，文法や SPOT との相関が相対的に高く，読解よりも作文との相関が高いことがわかる。判定結果が産出能力を示す作文との相関において高い数値を示していること，リスニング要素を含む SPOT との相関が高い（話す能力には聞く能力も関わる）ことから，本タスクは話す能力を測っており，基準関連妥当性（他のテストや外部基準との関係による妥当性）を一定程度満たしていると言えよう。

5. 考察

研究課題の信頼性については，判定者間の相関が高く，STAR は信頼性の高いテストだと言える。Kappa 係数は中程度の一致だったが，これは一部の発話に対する判定がずれた結果である。4 名が 6 つのレベルに判定した結果のうち，標準偏差 .8 以上の差が出た発話は 32 名中 7 名の発話だった。判定が割れたのは表 6 の理由だと考えられる。①以外はレベル 3 か 4 か 5 で判定が分かれていたことから中級の判定は初級や上級より難しいと言えそうである(5)。しかしながら，判定結果の順序には高い信頼性があった。

構成概念妥当性については，断りという 1 タスクではあるが，Celce-Murcia（2007）の包括的口頭コミュニケーション能力を概ねカバーしており，他のテストとの相関の結果，適切なレベルの基準関連もあり，概ね目標とする測定領域を測定していると判断できる。

実用性面では，録音機材さえあれば大勢が受験でき実施も簡便で短時間で初級から上級まで弁別できる。また，「断り」は現実の場面で行われる言語行動であり，本タスクは受

『日本語教育』175号（2020.4）
ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

表 6　判定が割れた代表例とその理由

	発話例	理由
①	クラスをやめてーちょっとー，えっ，とー，（沈黙）私は，えークラス，クラス，クラス，クラス，えー，行こうと，思っています。ごめん，ごめん，ごめん，ごめんね。えっとーまたーあとでーえークラブのことはー話してください。お先にー失礼します。	レベル 1 から 3 に判定が分かれた。途切れ途切れの話し方を文と判断したか単語と判断したか，断りはできているが配慮をどう判断したかで評価が分かれたと思われる。
②	ほんとにすみませんですけどー。なんかこの授業はほんとに楽しくてー，役にも立ってるんだからー，行きたいんだからー，あの，いまーすぐクラブに行くのをやめてー，あの，あとー，違う時にバンド聞きに行きませんか？誘ってくれてーほんとにうれし，ありがとうございました。	レベル 3 から 5 で判定が分かれた。完結している文が少ないこと，理由の述べ方の「んだから」を文法的間違いと判断したかで判定が分かれた可能性がある。
③	えー授業やめて？それは無理だよ無理です。私はの日本語はとても下手ですから。日本語はサボることはできないです。日本語はへ日本語は上手にならなきゃならない。はいそうです。えーごめんね？私はやっぱり日本語の授業の行きたいです。	レベル 3 から 5 で判定が分かれた。断りの配慮が少ないこと，また，状況説明がうまくないことから，評価が分かれた可能性がある。

験者である大学生が遭遇する場面として不自然でないか予め確かめた上で設定されたタスク（ボイクマン 2019）を元にしたことから，一定の真正性が確保されていると言えよう。

PT の全体構成から言えば，本テストを行う意義は大きい。実際，X 大学ではこの SP テストによって総合的に妥当なレベル認定のための情報が得られているだけでなく，書き言葉は苦手でも話し言葉は得意な受験者に適切な科目の受講を許可したりしている。

改善の余地は残るが，STAR は短時間でできる PT のための SP テストの役割を果たすことができ，音声サンプルとルーブリックによる判定方法には大きな可能性があると思われる。

6. まとめと今後の課題

状況対応の「断り」は，一貫性のある測定ができ信頼性が高く，包括的口頭コミュニケーション能力を測る妥当性のあるタスクであり，実用的で真正性もあることが検証できた。よって，STAR の「断り」タスクは PT に活用できる SP テストとしては十分な有用性があると考えられる。しかしながら，「断り」以外のタスクとの関連も今後検証する必要がある。そして，中級レベルの判定を厳密にするための判定ツールの改良も行いたい。

注

⑴ STAR は反応型テストだが，対面式でないため，「対話能力」を全面的には測れない。

⑵ 相互性とインパクトは，別途，実験や調査を行う必要があり，稿を譲りたい。

⑶ 検定には，R 2.8.1 を使用した。

⑷ 級内相関係数（ICC）（Shrout & Fleiss 1979）は，判定者間信頼性を表す指標の一つである。本研究のデータは順序尺度のため理論的にはカッパ係数でよいが，ICC は評価者，被評価者，交互作用，誤差の分散をすべて考慮に入れていて理論的に優れている（対馬 2002）ため，評点を間隔尺度とみなして，ずれの幅を調べるためにも ICC に意味があると考える。

⑸ 判定が分かれた要因を特定するため，判定者数を多くし，判定に迷った理由のコメント分

ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

析を行うなどのさらなる検証が必要だと考える。今後の課題としたい。

参考文献

⑴ 石塚賢吉・菊池賢一・篠崎隆宏・西村竜一・山田武志・今井新悟（2017）「日本語スピーキングテスト SJ-CAT の開発」https://www.slideshare.net/kenishiken/sjcat（最終アクセス 2019 年 4 月 15 日）

⑵ 国際交流基金（2017）『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（第二版第一刷改訂版）』独立行政法人国際交流基金

⑶ 小林典子・フォード丹羽順子・山元啓史（1996）「日本語能力の新しい測定法『SPOT』」『世界の日本語教育』第 6 号，201-236.

⑷ 対馬栄輝（2002）「理学療法の研究における信頼性係数の適用について」『理学療法科学』第 17 巻 3 号，181-187.

⑸ 波多野博顕（2018）「多相ラッシュ分析による JF 日本語教育スタンダード準拠ロールプレイテストの妥当性検証」『2018 年度 日本語教育学会秋季大会 予稿集』，161-166.

⑹ ボイクマン総子（2019）「日本語の『断り』における語用論的能力の発達」*Eruditi: The CGCS Journal of Language Research and Education*，vol.3, pp.14-28.

⑺ 牧野成一・鎌田修・齋藤真理子・荻原稚佳子・伊藤とく美・池崎美代子・中島和子（2001）『ACTFL-OPI 入門』アルク

⑻ Bachman, L. and Palmer, A. S.（1996）*Language testing in practice*. Oxford: Oxford University Press.

⑼ 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.* 41-57. Heidelberg, Germany: Springer Netherlands.

⑽ Council of Europe（2018）*CEFR Companion Volume with New Descriptors.* https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989（最終アクセス 2019 年 4 月 15 日）

⑾ JSST（Japanese Standard Speaking Test）https://www.alc.co.jp/jsst/ （最終アクセス 2019 年 4 月 15 日）

⑿ Shrout P.E. & Fleiss, J. L.（1979）. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86:420-428.

（東京大学）

ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会

Japanese Speaking Placement Test: Validation of a Task and an Evaluation Tool

BEUCKMANN Fusako, NEMOTO Aiko and MATSUSHITA Tatsuhiko

Ideally, a placement speaking test should be administered to as many examinees as possible within a short period of time; the evaluation procedure should be simple and convenient, with highly reliable and valid results. To accomplish this, a Japanese speaking test task with an evaluation tool consisting of rubric and voice samples was developed. Test reliability and validity were verified with the implementation of an experiment involving four Japanese teachers（raters）assessing the test results using the evaluation tool. The test task situation “refusal” was employed with 32 students from beginner to the advanced level participating. A high correlation between this speaking test and the writing placement test was observed, with a corresponding high correlation with the Simple Performance-Oriented Test（SPOT）which requires aural skills. In contrast, there was a low correlation with written receptive skills; regardless, the criterion-related validity of this test was reasonably and conclusively satisfied. The task and evaluation tools are well suited to evaluate the learners’ speaking ability, ensuring excellent reliability; assessing each learner’s level required only one to two minutes with high consistency and correlation among raters. Although intermediate level assessments appear to be more challenging, this test task and evaluation tool can be very valuable as a Japanese speaking test.

（The University of Tokyo）

『日本語教育』175号（2020.4）
ONLINE ISSN: 2424-2039
発行：公益社団法人日本語教育学会