

# スピーキングのレベル判定のための弁別性焦点化ルーブリック —非日本語教師による判定結果の分析—<sup>1</sup>

ボイクマン 総子<sup>2</sup>

根本 愛子<sup>3</sup>

松下 達彦<sup>4</sup>

DOI: 10.18999/stul.36.41

**要約:**本研究では、根本ほか(2020)で使用されたスピーキングテスト STAR (Speaking Test of Active Reaction) のルーブリックを改良するために、従前のルーブリック ver.1 と改訂版ルーブリック ver.2 を用い、非日本語教師 97 名による状況対応タスクのレベル判定比較実験を行った。ルーブリック ver.2 は、ver.1 の判定実験でのコメント分析結果をもとに、各評価項目をレベルごとに全て記述するのではなく、複数レベルにわたって記述することで、ルーブリックの記述の簡便化を図り、判定レベルの弁別性を焦点化した「弁別性焦点化ルーブリック」である。実験の結果、ver.2 は中級後半から上級レベルの弁別力が不足している可能性があるものの、ver.1 に比べ、受験者の音声を最後まで聞かなくても判定できる割合が上がった。また、レベル判定のための音声サンプルよりも ver.2 のルーブリックの方がわかりやすくなったとの結果が出た。このことから、ルーブリックを ver.2 にしたことで実用性が上がったと言える。本研究では、さらに、この弁別性を焦点化したルーブリックを状況対応タスクだけでなく、音読・シャドーイング・絵描写・再話・意見述べといった他のタスクにも応用し、これらのテストタスクのルーブリックを作成した。

**キーワード:** スピーキングテスト ルーブリック 弁別性 コメント分析 多相ラッシュ分析

---

<sup>1</sup> English title: Examining the Rubric Focusing on the Level Distinction for Japanese Speaking Test: Analysis of the Ratings by Non-Japanese Language Teachers

<sup>2</sup> BEUCKMANN, Fusako, The University of Tokyo, Japan, E-mail: beuckmann@g.ecc.u-tokyo.ac.jp

<sup>3</sup> NEMOTO, Aiko, The University of Tokyo, Japan, E-mail: anemoto@g.ecc.u-tokyo.ac.jp

<sup>4</sup> MATSUSHITA, Tatsuhiko, National Institute for Japanese Language and Linguistics, Japan, E-mail: tatsu.matsushita@ninja.ac.jp

## 1. はじめに

スピーキングやライティングなど産出型の言語テストにおいて、判定者が受験者のレベルを判定するツールとして、TOEFL や IELTS などでも見られるようにルーブリックが用いられることが多い。ルーブリックとは、「ある課題をいくつかの構成要素に分け、その要素ごとに評価基準を満たすレベルについて詳細に説明したもの」で、「課題・評価者尺度(達成レベル・成績評価点)・評価観点(課題が求める具体的なスキルや知識)・評価基準(具体的なフィードバックの内容)の全てを表形式で配置」したものである(スティーブンス・レビ, 2014)。ルーブリックは、大学の授業でレポートやプレゼンテーションなどの課題を評価する際に用いられることもあれば、言語テストのパフォーマンス評価として用いられることもある。

本研究では、根本ほか(2020)によってスピーキングテスト STAR(Speaking Test for Active Reaction)の状況対応タスクを判定するためのルーブリック ver.1 のコメント分析から得られた結果を踏まえ、修正した改訂版 ver.2 と ver.1 を比較し、ver.2 の信頼性と実用性を検証する。なお、ver.2 では、各評価項目をレベルごとに全て記述するのではなく、複数レベルにわたって記述し、判定レベルの弁別性を焦点化した。テスト開発では、判定基準は一度作成すればそれで完成ではなく、適宜、改良していくというプロセスをとることが望ましい。本研究は、テスト開発の一環として、より良い判定基準を作成するために行うものである。

## 2. 状況対応タスクのためのルーブリック ver.1 と ver.2

実験に用いられたテストタスクと判定ツールについて説明する。

ルーブリック ver.1 を用いた判定実験は、根本ほか(2020)の状況対応のテストタスクの受験結果をルーブリック ver.1 で判定したもので、その結果は、根本ほか(2020)にまとめられている。一方、ルーブリック ver.2 を用いた判定実験は、根本ほか(2020)の状況対応のテストタスクの受験結果を、新たに作成したルーブリック ver.2 で判定したもので、本研究では、両者の判定結果を比較検討する。

根本ほか(2020)の実験で用いられた状況対応タスクは、「断り」「勧誘」「依頼」のいずれかの言語行動が必要とされる状況に対応するもので、例えば、断りについては以下の状況(1)がコンピュータの画面上に示され、その後に出される音声キュー(2)に対して、1分以内で反応を録音するというものである。

(1)コンピュータ上に提示される状況説明文

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)日本語の音声キュー

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、授業行くのやめて、一緒に聞きに行かない？

表 1 STAR 状況対応判定用ルーブリック ver.1 (ボイクマンほか(2019), p.209)

構成要素	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がほとんどない	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

判定に用いられたツールは, ルーブリックおよび音声サンプルである。ルーブリック ver.1(表 1)は CEFR および OPI を参考に作成されたボイクマンほか(2019)で, 根本ほか

(2020)では、状況対応タスクを「ルーブリック ver.1」(表 1)と「音声サンプル」(表 2)を用いて判定している。音声サンプルは、過去の音声データからルーブリックにある各レベルの特徴が最も端的に表れているものをベンチマークとしてレベルごとに一つずつ選択されたものである。なお、STAR の判定レベルは、6 レベルとされており、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。

表 2 判定のための音声サンプルの文字化資料 (ボイクマンほか(2019), pp.209-210)

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せん、あの一あー、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、さんじゅうーにあります。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だけど、そのコンサートはた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。まなんか神社(?)さえも無理かなと思って。なんかあの一、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、はい。

根本ほか(2020)では、ルーブリック ver.1 を用いて、判定者に判定してもらったコメントを KH-Coder (樋口, 2020)を用いて分析した結果、

- ①「テキストの型」「対人配慮」「内容」では L3(中級前期)と L4(中級後期)を境界として、「文法」は L4(中級中期)と L5(中級後期)を境界として、レベル判定の観点が異なる。
- ②「自然さ」と「流暢さ」は L6(上級)の特徴である。
- ③中級前半は初級レベルと比較され肯定的なコメントが多い一方、中級後半は上級レベルと比較され否定的なコメントが多くなる。

ことを明らかにした上で、中級(L3～L5)の弁別力が十分でないことを指摘している。

そこで、本研究では、ルーブリックの構成要素はレベルにより特徴があるという特徴を活かし、弁別性を焦点化したルーブリック ver.2(表 3)を作成し、ver.1 と比較実験することで、ルーブリックの信頼性と実用性が改善されたか検証する。

表 3 STAR 状況対応判定用ルーブリック ver.2

構成要素	レベル1	レベル2	レベル3	レベル4	レベル5	レベル6
1) 流暢さ	(どちらかというと)流暢さが感じられない			(どちらかというと)流暢さが感じられる		非常に流暢
2)文法の正確さ	(どちらかというと)間違いが気になる				(どちらかというと)間違いが気にならない	
3) 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	(どちらかというと)段落レベルで話すことができる		
4) 内容の伝わり具合	(どちらかというと)課題達成に必要な内容が伝わらなかった			(どちらかというと)課題達成に必要な内容が伝わった		
5) 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
6) 対人配慮	配慮ができない、または、「すみません」や「ちよつと」を使った配慮ができる		最低限の配慮ができる	それなりの配慮ができる	十分な配慮ができる	

### 3. ルーブリック ver.1 と ver.2 を用いた判定実験の方法

判定実験で使用されたデータは、X 大学で 2018 年 9 月に実施されたスピーキングテストのうち、状況対応タスク 3 種(断り、勧誘、依頼)の 26 人分である。判定者は、ルーブリックと音声サンプル(表 2:各レベルを代表するベンチマークとなる音声サンプル1例ずつ×6 レベル)からなる「判定ツール」を用い、受験者の発話を 6 レベルに判定している。本実験では、この判定ツールのうち、用いるルーブリックのみを変え、ルーブリック ver.1 (表 1)を使ったグループ (以下、ver.1 群)と、ルーブリック ver.2(表 3)を使ったグループ(以下、ver.2 群)の判定結果を比較した。ver.1 群の実験は、2019 年 6 月から 7 月、ver.2 群の実験は、2021 年 10 月に行われた。

判定者となる実験協力者はいずれの群も非日本語教師であり、ver.1 群は 30 名、ver.2 群は 67 名である。ver.1 群は、状況対応タスクのうち「断りと勧誘」「断りと依頼」「勧誘と依頼」と 3 種類のうち 2 種類をランダムに振り分けられ、各判定者は 2 種類×学習者 26 名分=52 の

音声データを、ループリックと音声サンプルを用いてレベル判定を行った。一方, ver.2 群は, 3 種の状況対応のいずれかについて判定を行った。具体的には, 断りを判定したのは 20 名, 勧誘は 24 名, 依頼は 23 名であった。

両群の実験ともに, 判定者は, 音声サンプルを聞き, ループリックを読み, 十分に判定ツールを理解したと思ったら, 各自のタイミングで音声データを聞いて 1 から 6 のレベル判定を行なっている。なお, 判定に際し, 音声データをどの程度聞くかは, 判定者に任せた。

さらに, 判定者には, 判定レベルを選んだ後, コメント欄に「判定の際に迷ったこと, 気づいたこと, 判定根拠」などを自由に記入してもらった。これら一連のレベル判定実験の終了後に, 次の事後アンケート質問にも回答してもらった。質問項目は判定の方法に関わる内容で, (3)の 4 つの質問を 5 件法で行なった。

- (3) Q1 ループリックは役に立ったか。
- Q2 音声サンプルは役に立ったか。
- Q3 判定にどのくらい迷ったか。
- Q4 受験者の音声を最後まで聞いたか。

Q1と Q2 については, 「5=とても役に立った, 4=役に立った, 3=どちらともいえない, 2=役に立たなかった, 1=全然役に立たなかった」とした。Q3 は, 「5=ほとんど迷わなかった, 4=3 割ほど迷った, 3=半分, 2=7 割, 1=9 割以上迷った」とし, Q4 は, 「5=全部最後まで聞いた, 4=ほとんど最後まで聞いた, 3=最後まで聞く必要がなかった, 2=ほとんど最後まで聞かなかった, 1=全部最後まで聞かなかった」とした。

## 4. 結果

ver.1 群のデータは, 「受験者 26 名分の音声データ」×「1 タスク当たり 20 名の判定者(判定者総数 30 名)」×「状況対応 3 タスク」の全 1,560 の判定である。一方, ver.2 群のデータは, 「受験者 26 名分の音声データ×判定者 64 名×状況対応各 1 タスク」の全 1,664 の判定である。以下, 両群から得られたレベル判定結果, および, 実験後のアンケートの回答を比較する。

#### 4.1 ルーブリック ver.1 と ver.2 の多相ラッシュ分析

Facets Ver.3.83.0 (Linacre, 2020) を使用し, Linacre(1989) の多相ラッシュモデル (many-facet/multi-faceted Rasch measurement) に基づき, 「受験者」「評価者(本実験における「判定者」)」「タスク」の 3 つの相について, 当該テストがモデルに適合しているかを分析した。

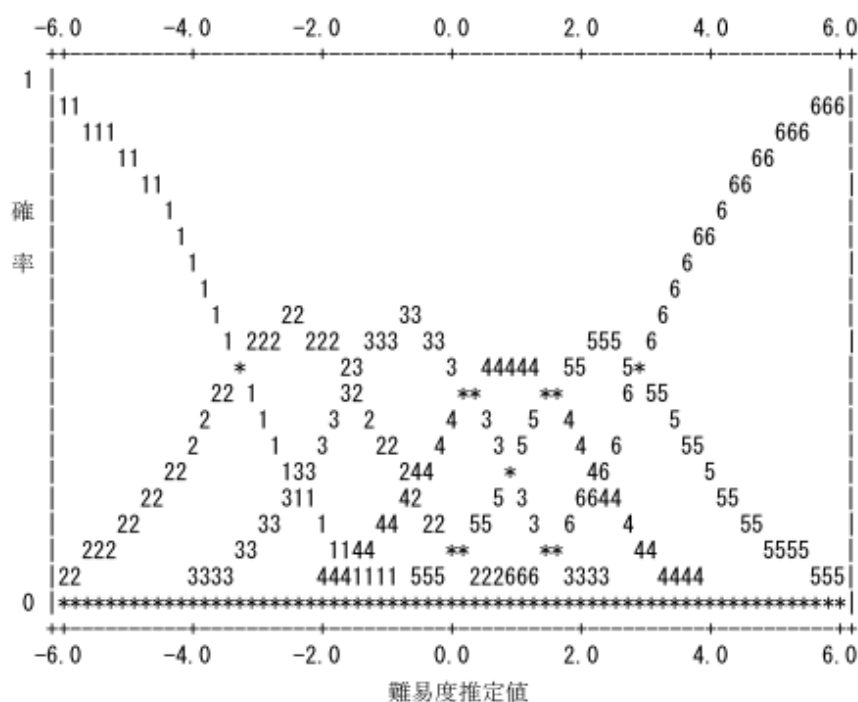


図1 ルーブリック ver.1 を使用した多相ラッシュ分析による難易度推定値の確率曲線  
(根本ほか, 2020)

ルーブリック ver.1 の多相ラッシュモデルによる分析は, 根本ほか(2020)が行なっている。それによると, ルーブリック ver.1 を用いた実験結果では, 「全体的にモデルによく適合し, 内的に一貫して評価されており, 概ねこの判定基準でこのタスクの評価ができ, タスクの質は良好で, 対象とする特性を測っている」との結果が出ている。さらに, 受験者分別信頼性, および, 判定者内信頼性についても, 高い一貫性があった。しかしながら, 確率曲線(図1)を見ると, 各レベルに概ねはっきりとした頂上があるが, 4 の頂上やや低く, 中級後半レベルで評価尺度の設定が十分ではないことが示唆された(根本ほか, 2020)。

では, 多相ラッシュモデルによる, ver.2 群の分析結果はどうか。ver.2 群では, 標準化誤差(Model. S.E.)の絶対値が 2 以上は 70 個で全体の 2.42%(70/2886)であり, 絶対値 3 以上

は 30 個で全体の 0.01%(30/2886)だった。標準化誤差は、通常、絶対値 3 以上が 1%程度、2 以上が 5%以下であれば良いとされており(Eckes, 2011), このデータの結果は問題ないことがわかる。モデル適合度については、インフィットの平均値(InfitMsSq)が 1.7 を超えている受験者が 3 名いたが、2.0 を超える者はいなかった。タスクの InfitMsSq は、2 以上が 5%以下であれば良いとされており(Eckes, 2011), このデータの結果は問題ないことがわかる。モデル適合度については、インフィットの平均値(InfitMsSq)が 1.7 を超えている受験者が 3 名いたが、2.0 を超える者はいなかった。タスクの InfitMsSq は、0.93~1.08 で、0.5~1.7 の範囲に収まっていた。判定者は、0.5 以下はいなかったが、1.7 以上が 2 名いた。データとモデルの予測が異なるケース(値が 1.7 以上)は一部に見られたが、大きく異なるケース(2.0 以上)はなかったため、モデルに概ね適合していると結論づけることができる。一次元性が満たされているかについては、点双列相関係数を見る必要がある。受験者、タスク、判定者の各相についてマイナスの値はなかったため、ver.2 群判定データの一次元性に問題はないといえる。以上のことから、ルーブリック ver.2 を使用した非日本語教師による判定データは、モデルに全体的にフィットしており、一次元性も満たされていることが確認できた。また、受験者分別信頼性は 1.00, 判定者内信頼性は .94 で、評価には一貫性もあることが検証された。

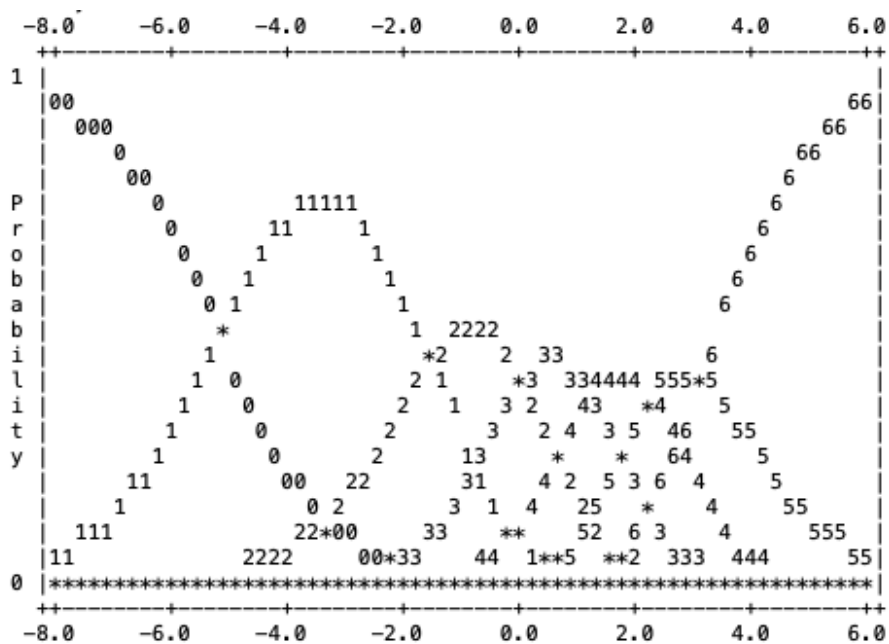


図 2 ルーブリック ver.2 を使用した多相ラッシュ分析による難易度推定値の確率曲線



しかしながら、図 2 に示すように、難易度推定値の確率曲線を見ると、各レベルに頂上は辛うじてあるが、レベル 4 と 5 の頂上が低く、ver.1 群と同様に、中級後半レベルで評価尺度の設定が十分ではないことが示唆された。

## 4.2 事後アンケート結果

事後アンケートでは、(3)の 4 つの質問を行った。質問を再掲する。実験後アンケートでは、ver.1 群からは 30 名、ver.2 群からは 64 名の回答が得られた(表 4)。

(3)Q1 ルーブリックは役に立ったか。

Q2 音声サンプルは役に立ったか。

Q3 判定にどのくらい迷ったか。

Q4 受験者の音声を最後まで聞いたか。

Q1 のルーブリックの役立ち度について、各群の解答の平均値に対して  $t$  検定を行ったところ、ver.1 群と ver.2 群では、役に立ったと回答した割合に有意差はなかった ( $t(92) = -1.329$ ,  $p = .434$ ,  $d = 0.762$ )。Q 2 の音声サンプルの役立ち度についても、両群に有意な差はなかった ( $t(92) = 3.827$ ,  $p = .943$ ,  $d = 0.677$ )。Q3 の迷い度についても同様に、有意差は見られなかった ( $t(92) = -1.298$ ,  $p = .281$ ,  $d = 0.954$ )。一方、Q4 の最後まで聞いたかとの質問については、ver.2 群の方が最後まで聞けなかったと答えた割合が有意に高かった ( $t(92) = -8.144$ ,  $p < .001$ ,  $d = 0.753$ )。なお、効果量も  $d = 0.753$  と十分ある。

表 4 事後アンケートの記述統計

	属性	度数	平均値	標準偏差
Q1 ルーブリック	ver.1 群	30	4.17	.913
	ver.2 群	64	4.13	.681
Q2 音声サンプル	ver.1 群	30	4.67	.661
	ver.2 群	64	4.09	.684
Q3 迷い度	ver.1 群	30	2.37	.850
	ver.2 群	64	2.64	.998
Q4 音声を最後まで聞いたか	ver.1 群	30	3.03	.928
	ver.2 群	64	4.39	.657

他に、ループリックか音声サンプルかどちらがわかりやすいかについては、ver.1 群は、音声サンプルの方が有意にわかりやすいと答えた( $t(58) = -2.430, p = .018, d = 0.797$ )。一方、ver.2 群は、ループリックの方が有意にわかりやすいと答えた( $t(126) = 2.461, p = .015, d = 0.682$ )。

### 4.3 コメント分析

コメント分析は、判定レベルを選んだ後、コメント欄に「判定の際に迷ったこと、気づいたこと、判定根拠」などを自由に記入してもらったものである。ver.1 群では、全 1,560 の判定のうち、レベル判定時のコメント記入欄に記入があったのは 1,038 件(66.5%)、ver.2 群は、全 1664 の判定のうち 1615 件(97.1%)であった。これらのコメントを KH Coder ver.3(樋口, 2020)で頻出語を判定レベル別に抽出した。

表 5 は、両群の上位 5 語を KH-Coder を用いて抽出した表である。表 5 から分かるように、ver.1 群ではコメントに表れた語彙がレベルによって異なるのに対し、ver.2 群では、レベル 1~4 までは上位語彙は、「伝わる」と「内容」であること、レベル 5~6 は、使用されている語彙が「流暢」と「表現」であるというように、違いが顕著に表れている。

表 5 判定レベル別頻出語上位 5 語の ver.1 群と ver.2 群の比較

ver.1 群						
全体	L1	L2	L3	L4	L5	L6
発音	単語	多い	多い	発音	発音	流暢
表現	多い	伝わる	内容	表現	表現	自然
流暢	文	表現	表現	流暢	少し	ネイティブ
多い	英語	内容	伝わる	少し	気	発音
自然	言う	文	発音	伝わる	流暢	日本人

ver.2 群						
全体	L1	L2	L3	L4	L5	L6
発音	伝わる	伝わる	伝わる	伝わる	流暢	流暢
表現	内容	内容	内容	内容	表現	表現
流暢	単語	流暢	流暢	表現	伝わる	配慮
多い	伝える	表現	表現	流暢	内容	自然
自然	レベル	文法	文法	配慮	配慮	豊か

## 5. 考察

ver.1 群と ver.2 群の比較実験の結果、ルーブリックを全てのレベルの項目について記述するのではなく、項目ごとの特徴を考慮し、複数レベルにわたって記述することで、判定レベルの弁別性を焦点化した ver.2 の方が、わかりやすいという結果が得られた。コメント分析の結果からは、ver.2 群の方が、何をもとに判定をしているのかがレベルごとに明確化されており、項目ごとに焦点となるレベルが異なっていることから、ver.2 群の方がレベルの特徴を捉えて判定を行なっていることが示唆された。

判定にかかる時間については、ver.2 の方が最後まで聞かなくても自信を持って判定できており、時間短縮という効率化に貢献していることがわかった。

多相ラッシュ分析の結果からは、両群ともに、全体的にモデルに適合しており、内的にも一貫して評価されていて、どちらの判定基準でもタスクの評価ができていたことがわかった。しかしながら、両群とも、中級後半レベルの判定で、差別化ができておらず、さらなる改良の余地があることが明らかになった。

## 6. 他のタスクのルーブリック

本研究の実験結果により、ver.2 のように弁別性を焦点化したルーブリックの有用性が明らかになったことから、ここでは、STAR の他のタスクについても、弁別性を焦点化したルーブリックを提案する。表6は音読タスク、表7はシャドーイングタスク、表8は絵描写、表9は再話、表10は状況対応、表11は意見述べのルーブリックである。

英語のシャドーイングに関する測定法に「音節法」による測定法がある(門田, 2018)。これは、正しく再現できた音節の数を数える方法で、本研究ではこれを応用し、どの程度再現できていたかをその基準の1つとした。音読タスクでは、「正読率」を設け、書かれてある内容をどの程度再現できたか(表6)、シャドーイングタスクでは、「正聞話率」を設け(表7)、聞いた内容をどの程度再現できていたかを判定者の印象で判断させることにする。

表 6 音読タスクのルーブリック

構成要素	レベル 1	レベル 2	レベル 3	レベル 4	レベル 5	レベル 6
流暢さ(リズム,イントネーション,スピード)	(どちらかというと)流暢さが感じられない			(どちらかというと)流暢さが感じられる		非常に流暢
正読率	10%以下	30%程度	50%程度	70%程度	85%程度	ほぼ 100%

表 7 シャドーイングタスクのルーブリック

構成要素	レベル 1	レベル 2	レベル 3	レベル 4	レベル 5	レベル 6
流暢さ(リズム,イントネーション,スピード)	(どちらかというと)流暢さが感じられない			(どちらかというと)流暢さが感じられる		非常に流暢
正聞話率	10%以下	30%程度	50%程度	70%程度	85%程度	ほぼ 100%

絵描写タスク(表 8)は, 状況対応タスクのルーブリック ver.2 から「対人配慮」を省き, 第 4 項目の「話の伝わり具合」を「話の内容の再現度」に変えた。絵を見て描写するタスクは, 絵に描かれてある状況をどの程度再現できているかがレベル判定の鍵となる。

表 8 絵描写タスクのルーブリック

構成要素	レベル 1	レベル 2	レベル 3	レベル 4	レベル 5	レベル 6
1) 流暢さ	(どちらかというと)流暢さが感じられない			(どちらかというと)流暢さが感じられる		非常に流暢
2) 文法の正確さ	(どちらかというと)間違いが気になる				(どちらかというと)間違いが気にならない	
3) 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	(どちらかというと)段落レベルで話することができる		
4) 話の内容の再現度	話の内容が説明できていない	一部の話の内容のみ説明できている箇所がある	情報に漏れがあり流れの断絶が見られるが, 部分的に説明できている	情報に漏れがあり流れの断絶も多少見られるが, 概要は説明できている	詳細部分の抜けがあるが, 概ね説明できている	詳細な点まで説明できている
5) 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	慣用句や比喻なども交えて話することができる

再話タスク(表 9)は, 絵描写タスクと同様に, 状況対応タスクのルーブリック ver.2 から「対人配慮」を省き, 4つ目の項目の「話の伝わり具合」を「話の内容の再現度」に変えた。話を

聞いてそれと同じ内容を再度伝える再話タスクは、話された内容がどの程度再現できているかがレベル判定の鍵となる。

表9 再話タスクのルーブリック

構成要素	レベル1	レベル2	レベル3	レベル4	レベル5	レベル6
1) 流暢さ	(どちらかという)流暢さを感じられない			(どちらかという)流暢さを感じられる		非常に流暢
2) 文法の正確さ	(どちらかという)間違いが気になる				(どちらかという)間違いが気にならない	
3) 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	(どちらかという)段落レベルで話すことができる		
4) 話の内容の再現度	話の内容が再現できていない	一部の話の内容のみ再現できている箇所がある	情報に漏れがあり流れの断絶が見られるが、部分的に再現できている	情報に漏れがあり流れの断絶も多少見られるが、概要は再現できている	詳細部分の抜けがあるが概ね再現できている	詳細な点まで再現できている
5) 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

状況対応タスクのルーブリック(表10)は、ver.2(表3)を再掲したものである。

表 10 状況対応判定用ルーブリック (表 3 再掲)

構成要素	レベル1	レベル2	レベル3	レベル4	レベル5	レベル6
1) 流暢さ	(どちらかという)流暢さが感じられない			(どちらかという)流暢さが感じられる		非常に流暢
2)文法の正確さ	(どちらかという)間違いが気になる				(どちらかという)間違いが気にならない	
3) 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	(どちらかという)段落レベルで話することができる		
4) 内容の伝わり具合	(どちらかという)課題達成に必要な内容が伝わらなかった			(どちらかという)課題達成に必要な内容が伝わった		
5) 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	慣用句や比喻なども交えて話すことができる
6) 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		最低限の配慮ができる	それなりの配慮ができる	十分な配慮ができる	

意見述ベタスクのルーブリック(表 11)は、状況対応タスクのルーブリック ver.2 から「対人配慮」を省き、4つ目の項目の「話の伝わり具合」を「論理性」に変えたものである。意見を述べる際には、論理的に述べられているかどうかを重要となると考えたからである

表 11 意見述ベタスクのルーブリック

構成要素	レベル1	レベル2	レベル3	レベル4	レベル5	レベル6
1) 流暢さ	(どちらかという)流暢さが感じられない			(どちらかという)流暢さが感じられる		非常に流暢
2)文法の正確さ	(どちらかという)間違いが気になる				(どちらかという)間違いが気にならない	
3) 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	(どちらかという)段落レベルで話することができる		
4) 論理性	意見が述べられないか、ごく簡単な意見のみ述べられる		意見は述べられているが、根拠や理由の説明が不十分である	十分な理由や根拠に基づいて、意見が述べられている		
5) 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	慣用句や比喻なども交えて話すことができる

## 7. おわりに

本研究では、各項目をレベルごとに全て記述したルーブリック ver.1 と弁別性を焦点化したルーブリック ver.2 を用いて、非日本語教師 97 名による状況対応タスクのレベル判定比較実験を行った。その結果、両バージョンとも、中級後半から上級レベルの弁別力が不足している可能性があるものの、ver.1 に比べ、受験者の音声を最後まで聞かなくても判定できる割合が上がり、時短という実用性が上がった。

そして、本研究では、さらに、この弁別性を焦点化したルーブリックを状況対応タスクだけでなく、音読・シャドーイング・絵描写・再話・意見述べといった他のタスクにも応用し、これらのテストタスクの弁別性焦点化ルーブリックを作成した。

今後の課題としては、中級後半から上級レベルの弁別性を高めるため、このレベルの判定に迷うケースを用いて、判定者トレーニングを行うことである。また、状況対応以外のタスクのルーブリックの検証も行いたい。

## 付記

本研究は、JSPS 科研費 20K00697 (研究代表者ボイクマン総子)、および、20KK005 (研究代表者ボイクマン総子)の助成を受けている。

## [参考文献]

- 門田修平(2018)『外国語を話せるようになるしくみ シャドーイングが言語習得を促進するメカニズム』, SB クリエイティブ.
- スティーブンス, ダネル・レビ, アントニア(2014)『大学教員のためのルーブリック評価入門』佐藤浩章(監訳), 玉川大学出版部.
- 根本愛子・ボイクマン総子・松下達彦(2020)「状況対応タスクの非日本語教師による判定の分析—プレースメントのための日本語スピーキングテストの検証—」『日本語教育』177 号, 1-16.
- 樋口耕一(2020)『社会調査のための計量テキスト分析 —内容分析の継承と発展を目指して— 第2版』, ナカニシヤ出版.
- ボイクマン総子・根本愛子・松下達彦(2019)「同一の判定基準を用いて異なる言語行動を判

定することが可能か—プレースメントのための日本語スピーキングテスト STAR の判定基準の検証—」『2019 年度日本語教育学会春季大会予稿集』, 208-213.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang Pub Inc.

Frankfurt am Main: Peter Lang. Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

.....

ボイクマン 総子

(東京大学大学院総合文化研究科・教授)

Email: [beuckmann@g.ecc.u-tokyo.ac.jp](mailto:beuckmann@g.ecc.u-tokyo.ac.jp)

根本 愛子

(東京大学大学院総合文化研究科・准教授)

Email: [anemoto@g.ecc.u-tokyo.ac.jp](mailto:anemoto@g.ecc.u-tokyo.ac.jp)

松下 達彦

(国立国語研究所・教授)

Email: [tatsu.matsushita@ninjal.ac.jp](mailto:tatsu.matsushita@ninjal.ac.jp)