

令和 5 年 6 月 17 日現在

機関番号：12601

研究種目：基盤研究(C)（一般）

研究期間：2020～2022

課題番号：20K00697

研究課題名（和文）プレースメントのための日本語スピーキングテスト（STAR）の開発と検証

研究課題名（英文）Development and Validation of Japanese Speaking Test (STAR) for Placement

研究代表者

ボイクマン 総子（梶本総子）（Beuckmann, Fusako）

東京大学・大学院総合文化研究科・教授

研究者番号：50370995

交付決定額（研究期間全体）：（直接経費） 2,800,000円

研究成果の概要（和文）：本研究の目的は、プレースメントのための日本語スピーキングテスト STAR (Speaking Test of Active Reaction) を開発し、その有用性、信頼性、構成概念妥当性、真正性、相互性、インパクト、実用性の検証を行うことである。本研究では、簡便に実施できるSTARのテストタスクを主に状況対応タスク(ロールプレイタスク)を中心に開発し、経験のある日本語教師だけでなく、日本語母語話者の非日本語教師でも短時間のトレーニングで信頼性の高い判定をすることが可能なテストにするため、判定ツールであるルーブリックを改良した。

研究成果の学術的意義や社会的意義

既存の日本語スピーキングテストは、主に実用性の点においてプレースメントテスト向きではなく、統計的な検証も十分に行われているとは言い難いため、プレースメントテストのためのスピーキングテストを本研究で開発し、その有用性を検証することは、日本語教育においても、また、日本語スピーキングテスト開発の点からも有用であると考えられる。

本研究が解明しようとする学術的な問いのはプレースメントテストにふさわしい日本語のスピーキングテストとはどのようなテストかである。

研究成果の概要（英文）：The purpose of this study is to develop the STAR (Speaking Test of Active Reaction) Japanese speaking test for placement, and to verify its usefulness, reliability, construct validity, authenticity, interactivity, impact, and practicality. In this study, we developed STAR test tasks mainly focusing on situational response tasks (Role-play tasks) that can be implemented conveniently. We aimed to create a test that not only experienced Japanese teachers but also non-Japanese teachers who are native speakers of Japanese can make reliable judgments with a short period of training. To actualize this aim, we have refined the rubric, which serves as an assessment tool.

研究分野：言語学

キーワード：言語テスト スピーキングテスト スピーキング能力 第二言語習得 妥当性 信頼性 多相ラッシュモデル コメント分析

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

1. 研究開始当初の背景

日本語学習者の場合、話す能力とそれ以外の能力に差がある者も少なくない。そのため、筆記試験だけでは適切なレベルや科目に学生を配置するのは難しく、プレースメントテスト(PT)でスピーキングテストを実施することが重要である。また、PTにおいては、来日前であっても受験できるよう、世界各国からアクセスできるオンラインによるテストの実施が望ましい。

しかしながら、大勢が短時間で受験でき、実施と判定が簡便で、判定者間で一貫性の高い結果が得られるようなPT用の日本語スピーキングテストは研究開始当初は管見の限りなかった。また、オンラインで実施できるスピーキングテストはあるが、受験時間、判定者のトレーニングの点から、PTとしての実用性の観点に難がある。

このように、既存の日本語スピーキングテストは、総じてPT向きではなく、また、どれも統計的な検証が十分に行われているとは言い難いのが現状であった。

2. 研究の目的

本研究の目的は、日本語教育に特別の知識のない日本語母語話者でも、また、どの日本語教育機関でも簡便に行えるPTのための日本語スピーキングテストSTAR(Speaking Test of Active Reaction)を開発し、その有用性の検証を実証的に行うことである。

STARはループリックと音声サンプルからなる判定ツールを用いて初級から上級までの、CEFRに対応する6レベルに判定できる、オンラインによる半直接形式のスピーキングテストである。テストタスクは、スピーキング能力全般を測れるようにするため、多様なタスクを開発するのが望ましいと考え、(1)音読、(2)シャドーイング、(3)絵を見て描写、(4)話を聞いて再話、(5)状況対応、(6)意見述べの6種類を用意し、それぞれのタスクの判定ツールを開発・検証することを研究の目的とした。

また、本研究では、日本語教育に特別の知識のない日本語母語話者や経験の浅い日本語教師でも簡便に判定ができる判定ツールを開発するために、6種類のタスクのうち、特に、(5)状況対応タスクに焦点をあて、日本語教師と非教師(日本語母語話者)の判定の方法を比較し、両者の違いを判定ツールに反映し、改良することも行なった。

3. 研究の方法

本研究では、まず、状況対応タスク(断り)を用いて予備実験を行い、本タスクと判定ツールの信頼性と妥当性を検証した。実験協力者は、初級から上級までの受験者32名、判定者は日本語教師4名による。この判定結果と、他の科目のPT(漢字、文法、語彙、作文、SPOT)とを比較し、本テストタスクが基準関連妥当性を満たしているか、および、信頼性があるか、検証した(発表論文)。

本調査として、状況対応タスク(断り、勧誘、依頼)の受験結果について、非日本語教師30によるレベル判定実験を行なった。なお、対象としたテスト結果は、初級から上級までの26名であった。判定者によるレベル判定の根拠を得るため、判定結果に関するコメント記述を行ってもらい分析データとした(発表論文)。

日本語教師と非日本語教師の当該テストの判定基準の比較を行うため、10年以上の日本語教師経験のある教師33名、教師経験年数が3年から10年の日本語教師15名、非日本語教師64名の計112名を対象に、状況対応タスク(断り、勧誘、依頼)における判定方法と判定観点の異なりを検証した。対象としたテスト結果は、初級から上級までの26名である(発表論文)。分析には、多相ラッシュモデルのFACETによる分析とKH Coderによるコメント分析を行なった。

以上、主として、状況対応のタスクを実験対象タスクとして検証を行なった。

4. 研究成果

3つの発表論文を中心に、それぞれの研究成果を述べる。

パイロット研究(発表論文)から明らかになったことは以下の通りである。STARは、プレースメント時の読解などの受容能力よりも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ことと言える。判定者間の一貫性と相関が高く、受験者一人当たりの判定時間が1~2分であったことから、実験に用いたタスクと判定ツールである、音声サンプルと状況対応のループリックver.1は、簡便で信頼性も確保できていることがわかった。しかし、標準偏差.8以上の差が出た発話が一部にあり、それはレベル3か4、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に比べ難しいことがわかった。構成概念妥当性については、包括的口頭コミュニケーション能力(Celce-Murcia, 2007)を概ねカバーしていることが明らかとなった。実用性の面からも、大勢が受験でき実施も簡便で短時間で初級から上級の6レベルまで弁別できることがわかった。

上述のパイロット調査をもとにした本調査(発表論文)の結果、明らかになったことは次の通りである。まず、判定結果を、Facets Ver.3.83.2(Linacre, 2020)を使用し、Linacre(1989)

の多相ラッシュモデルに基づいて「受験者」「評価者(判定者)」「タスク」の3つの相について分析した。その結果、標準化残差(StRes)は絶対値2以上の場合が5%、絶対値3以上が1.2%でモデルに適合していることがわかった。infitとoutfitのMnSqの値が0.5-1.7の範囲を外れたmisfitは受験者1名と判定者2名で、他は全て範囲内であったことから、全体的にモデルによく適合しており、内的に一貫して評価されており、概ね本判定基準(ルーブリック ver.1)でこのタスクの評価ができていて、また、タスクの質も良好で、対象とする特性を計っていることが検証できた。受験者分別信頼性は0.99で、評価には高い内的一貫性があることも示された。さらに、「断りと勧誘」「断りと依頼」「勧誘と依頼」のグループごとに級内相関係数を計算したところ、判定者ない信頼性の指標であるICC(1,1)はそれぞれ、.814、.748、.844であった。絶対一致に基づくICC(3,1)と相対一致に一致に基づくICC(2,1)は、それぞれ、.815、.750、.844と.835、.813、.861で、いずれも幸運に高かった。しかしながら、タスクの相については、4.54で、設定した判定段階数の6に満たなかったため、ある程度はてんでいできてはいるが、十分には分別しきれない可能性が示唆された。これらのことから、STARの状況対応タスクはいずれも信頼性が高いと言えるが、発表論文でも指摘されているように、中級レベルの判定は初級や上級より難しく、中級を中心に判定ツールの精緻化を行うなど、判定の絶対一致を高める工夫が必要であることも明らかとなった。

発表論文では、判定時に記入してもらったコメントも分析した。全コメント数、1038件をKH Coder ver.3(樋口, 2014)で頻出ごとその使用分容を判定レベル別に分析した。その結果、判定レベルが上がるとコメント数が増える上昇型、最初はコメント数が増加するものの一定のレベルを超えるとコメント数が減少する山型の構成要素があることがわかった。上昇型のコメントには「配慮」「正確さ」「流暢さ」があった。「配慮」はレベル1(L1)とL2ではほとんどコメントがなく、L3で肯定と否定のコメントが増加、L4とL5では肯定的評価が増加すると同時に否定的評価は減少、L6では否定的なコメントは見られなかった。このことから、評価項目である「対人配慮」は、L3とL4を境に評価の観点が変化し、すなわち、能力がこのレベルを境に変化することがわかった。「正確さ」と「流暢さ」についても、L1とL2はコメントはほとんどなく、あっても否定的なコメントであった。L3から4にかけて否定と肯定のコメントが増加し、L4とL5では肯定的評価が増加、L6では否定的コメントが見られないという傾向が見られた。他方、山型をとる構成要素を見てみると、「課題達成」と「テキストの型」はL3からL4で、「文法」はL4からL5で、「表現」はL5からL6の間でそれぞれコメントが減少しており、これらのレベルを境にレベル判定が不要であると判断された可能性が示唆された。「自然/不自然」というコメントは、L5までは斬増するが、L6で肯定的コメントが急増することから「自然さ」は、L6の特徴であることがわかった。「内容」については、L3からL4で急激に減少した後で増加しており、L3とL4の間で判定の観点が変化したと考えられる。「伝わる/伝える」については、L4をピークに徐々に減少する一方で、L4以降は「伝わらなかった」というコメントは見られなかったことから、タスク完遂のための意図が伝わったか否かの判断は、L3とL4で差があると考えられる。以上、コメントをルーブリックと照合したところ、「テキストの型」「対人配慮」「内容」では、L3(中級前期)とL4(中級中期)で、「文法」は、L4(中級中期)とL5(中級後期)の間でレベル判定の観点が異なることがわかった。さらに、「自然さ」と「流暢さ」はL6(上級)の特徴であることも明らかになった。そして、L3(中級前期)は初級レベルのL1やL2と比較され肯定的なコメントが多い一方、L4(中級中期)は、L5やL6と比較され、否定的なコメントが多くなることもわかった。このように判定時のコメント分析から、ルーブリックの構成要素は、レベルにより特徴があると言え、この特徴を活かしてルーブリックを改善する示唆が得られた。この示唆を利用して、ルーブリックの改良(ver.2)を行い、発表論文で公表した。

発表論文では、教師と非教師とで信頼性に差があるか、教師経験の違いによってテスト判定方法と判定の観点が異なるかを検討した。その結果、教師と非教師でテスト判定の信頼性に大差は見られなかったものの、非教師の方が中級から上級レベルの弁別に難があることが、Facetsの分析結果から明らかとなった。なお、発表論文では、状況対応ルーブリック ver.2を用いた。

教師経験の違いでテストの判定の仕方が異なるかについては、事後アンケートとして、質問紙による4つの質問 1)ルーブリックは役に立ったか、2)音声サンプルは役に立ったか、3)判定にどのくらい迷ったか、4)受験者の音声を最後まで聞いたか を行い、5件法で分析した。分析の結果、10年以上の経験のある教師は非教師よりも音声サンプルを重視する傾向があることがわかった。また、非教師は、レベルによって判定の観点を変えることがないのに対して、教師は判定の観点を変えていた。具体的には、初級前半(L1かL2かの判定)は「流暢さ」、初級後半(L2かL3かの判定)は、「テキストの型」、中級前期と中期(L3かL4かL5かの判定)は「表現の豊かさ」、中級後期と上級(L5かL6かの判定)は「対人配慮」と「流暢さ」によって判定を行っていたことが明らかとなった。しかも、この傾向は、10年以上の経験のある教師により強く見られた。この結果により、STARは、非教師である日本語母語話者でもある程度の判定はできるものの、日本語教育経験の長い教師に対してよりも、より入念なトレーニングを実施する必要があることがわかった。そして、その際には、上述したベンチマークとなる判定基準を判定者トレーニングで強調すると、よりよりトレーニングが行えるであろうということが示唆された。

状況対応タスク以外のタスクについては、学会発表で、音読タスクとシャドーイングタスク

の成果について発表を行なった。さらに、学会発表 で、意見述ベタスクについての成果発表を行なった。これらの成果は今度、発表論文の形で公表を行う予定である。

本研究では、さらに、オンラインで受験できる受験サイトと、オンラインで判定できる判定者サイトを構築した。

(<https://speaking.komabastar.com/login>)

これらのサイトを通して、受験者は6つのタスクについてオンライン受験ができ、これまでに約200名の受験者によって利用された。また、研究発表 の判定実験は、この判定サイトを利用して行われた。

今後、タスクの種類と設問の検証を重ね、一般公開を行う予定である。

5. 主な発表論文等

〔雑誌論文〕 計4件（うち査読付論文 3件/うち国際共著 0件/うちオープンアクセス 4件）

1. 著者名 ボイクマン総子・根本愛子・松下達彦	4. 巻 24巻2号
2. 論文標題 スピーキングテスト判定における日本語教師と非日本語教師の比較-状況対応タスクに対する判定結果-	5. 発行年 2022年
3. 雑誌名 社会言語科学	6. 最初と最後の頁 35-48
掲載論文のDOI（デジタルオブジェクト識別子） 10.19024/jajls.24.2_37	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 ボイクマン総子・根本愛子・松下達彦	4. 巻 175
2. 論文標題 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-	5. 発行年 2020年
3. 雑誌名 日本語教育	6. 最初と最後の頁 146-154
掲載論文のDOI（デジタルオブジェクト識別子） 10.20721/nihongokyoiku.175.0_146	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 根本愛子・ボイクマン総子・松下達彦	4. 巻 177
2. 論文標題 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	5. 発行年 2020年
3. 雑誌名 日本語教育	6. 最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.20721/nihongokyoiku.177.0_1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 ボイクマン総子・根本愛子・松下達彦	4. 巻 36
2. 論文標題 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	5. 発行年 2022年
3. 雑誌名 ことばの科	6. 最初と最後の頁 41-56
掲載論文のDOI（デジタルオブジェクト識別子） 10.18999/stul.36.41	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

〔学会発表〕 計4件（うち招待講演 0件 / うち国際学会 0件）

1. 発表者名 ボイクマン総子・根本愛子・松下達彦
2. 発表標題 日本語のスピーキングテストにおける音読およびシャドーイングタスクの開発と妥当性検証 「論証に基づく妥当性検証の枠組み」を用いて
3. 学会等名 日本言語テスト学会
4. 発表年 2021年

1. 発表者名 ボイクマン総子・根本愛子・松下達彦
2. 発表標題 論証に基づく妥当性検証の枠組みによるスピーキングテストの検証 意見述べてを例として
3. 学会等名 日本語教育学会
4. 発表年 2022年

1. 発表者名 峯松信明・梶原卓弥・斉藤一弥・ボイクマン総子
2. 発表標題 L2英語及び日本語音声を対象とした自動採点とその精度
3. 学会等名 言語教育 エキスポ 2020、大学英語教育学会(JACET)
4. 発表年 2021年

1. 発表者名 梶原卓弥・瀧陽・斎藤大輔・峯松信明・斉藤一弥・ボイクマン総子
2. 発表標題 英語学習者と日本語学習者音声に対する多角的評価に関する実験的検討
3. 学会等名 日本音響学会、2021年春季研究発表会
4. 発表年 2021年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究 分担者	根本 愛子 (NEMOTO AIKO) (80746701)	東京大学・大学院総合文化研究科・准教授 (12601)	
研究 分担者	松下 達彦 (MATSUSHITA TATSUHIKO) (00255259)	国立国語研究所・大学共同利用機関法人人間文化研究機構・教授 (62618)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------