様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １４ 日現在

機関番号：３２６２０
研究種目：基盤研究(C)（一般）
研究期間：2014～2018
課題番号：２６３７０７３７

研究課題名（和文）スピーキング能力測定に向けたロールプレイテストの開発と妥当性検証

研究課題名（英文）Development and validation of a role-play test to evaluate English speaking ability

研究代表者
小泉　利恵（KOIZUMI, Rie）

順天堂大学・医学部・准教授

研究者番号：７０４３３５７１

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：　本研究の目的は、英語授業中に実施可能なロールプレイテストを開発し、妥当性を確保しつつ実施する方法を確立することである。ロールプレイテストのより適切な実施・採点方法を、ディスカッションテストと比較しながら調べ、実施可能なテスト方法を提案した。日本人大学生を対象にテストを実施し、Chapelle, Enright, and Jamieson (2008) に基づいて多くの妥当性の側面から分析を行い、ロールプレイテストがテストとして適切な性質を持ち、「提案する」「感謝・謝罪する」などの発話機能を多く引き出せたことなど、利点を明確にすることができた。

研究成果の学術的意義や社会的意義
包括的な検証がなされていなかったロールプレイテストの特徴を調べることで、言語テスティング研究への知見の提供につながった。また、英語スピーキング能力の指導の成果を測る1方法として、経験が少ないと実施しにくいロールプレイテストの手順を明確にし、評価方法を明示することにより、スピーキング評価の促進や、コミュニケーション能力の評価の改善に貢献できた。
　さらに、テストの妥当性検証方法を英語教育学に紹介できた。本研究が一適用例として、妥当性検証の手法を紹介することで、今後のテスト開発に影響を与えることができた。

研究成果の概要（英文）：This study aims to develop a role-play test for students of English as a second language and establish test procedures while maintaining the validity of interpretations and uses based on the test scores. We compared the role-play test with a discussion test to develop effective test administration and evaluation methods. Japanese undergraduate students of English took the role-play and discussion tests, and we analyzed their test performance data and responses to a questionnaire regarding various validity aspects based on Chapelle, Enright, and Jamieson (2008). The results indicate that the role-play test has multiple appropriate properties and advantages such as the ability to elicit various language functions including “proposing” and “thanking and apologizing.”

研究分野：英語教育学

キーワード：言語テスティング　英語テスト　スピーキング能力測定　ロールプレイテスト　ディスカッションテスト　ペア型対話テスト

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

英語スピーキング能力の育成のための指導には多くの時間が費やされているが、その評価は十分に行われていない。本研究では、ロールプレイテストのより適切な実施・採点方法を調べ、授業中に実施可能な評価方法を提案する。ロールプレイテストとは、学生同士で、カードで指定した役割で会話する形式で、その採点の困難さから、授業での評価時にはあまり使われていない。

英語教育では、英語コミュニケーション能力、特に他者と英語で対話・やり取りができる能力の育成が求められている（例：文部科学省, 2009)。能力育成のためには、効果的な指導を行うだけでなく、定期的に評価を行い、その結果に基づき、弱点を補強し、長所を伸ばしていくことが必要である。しかし、日本では教室でのスピーキング能力評価はまだ十分に行われていない。その評価を普及するには、教師の知識や意識を高めるだけでなく、教師の使用しやすい評価方法を開発することが必要である。

スピーキングテスト形式には、モノローグ型（例：スピーチ）と対話型（例：面接）があり、多面的な測定のためには両方が求められる。対話型テストではモノローグ型よりも多くの要素が測りたい能力以外に影響する傾向があるが、モノローグ型の研究ほど研究が進んでおらず、評価の実践にも生かすことが難しい状況である。これは、対話力の育成を重視する教育の中で大きな欠落であり、ロールプレイテスト開発により、その状況を改善することを目指す。

対話型テストの中の、学習者同士が話すロールプレイテストは有用であるが、その実証研究は限られる。評価を行う際に考慮すべき点には多様な要素があり（Fulcher, 2003)、先行研究もある（例：Csépes, 2009; Ducasse, 2010; Teng, 2013)。しかし、ロールプレイに焦点を当てた詳細な研究は知る限りない。本研究は、関連要素の多くの側面から分析する。

２．研究の目的

英語授業中に実施可能なロールプレイテストを開発し、妥当性を確保しつつ実施する方法を確立することが目的である。本研究では、2 名の学生同士が、カードで提示された役割を演じながら対話する形式をとるスピーキングテストを、ロールプレイテストと呼ぶ。本テストは採点の困難さから、授業での評価時にはあまり使われていない。本研究は、日本人大学生を対象とし、Chapelle, Enright, and Jamieson (2008) に基づき、多くの妥当性の側面から分析し、ロールプレイテストの妥当性検証として提示すること、またロールプレイテストの特徴を示すことを目的とする。

３．研究の方法

①研究１（図書④）

ロールプレイテストと、比較対象とするディスカッションテストの手順を開発し、4 タスクと１種類のルーブリック（評価基準）を作成し、日本人大学生 163 名に対して予備研究を行った。タスク作成の際には、O’Sullivan, Weir, and Saville (2002) のスピーキングの言語機能を参照し、学生同士の対話に特徴的な発話が現れるようなタスクにすることを意識した。ルーブリックは、Nakatsuhara (2013), Taylor (2011) 等を参考に、まずは 3 段階の総合的尺度を作成した。結果に基づき、形式やタスクを改善した。

②研究２（図書③;　テストの紹介を図書②として発行）

修正後のロールプレイテストとディスカッションテストを実施し、発話を録音した。対象は 190 名の日本人大学生であり、11 個のタスクを行った。多相ラッシュ分析、一般化可能性理論、確認的因子分析を行いて、テストとして望ましい性質があるかを確認した。使用したロールプレイとディスカッションのタスクの例を表１に挙げる。

表１．タスクの例

ロールプレイ	ディスカッション
Card A あなた（Student A）は虫歯（toothache）が痛くて困っています。ひどい状況を Student B に伝えてください。B のアドバイスを、理由を言って、1 回は否定してください。しかし、可能そうな案が出てきたら、今後どうするかを伝えてください。その後も会話を続けます。	**Card A, Card B 共通** Hobby（例：sports, club, last weekend, Golden Week)
Card B Student A は虫歯が痛くて困っています。状況を理解してあげてください。あなた（Student B）は、歯医者（dentist）に行くことや痛み止め（painkiller）を飲むことを勧めます。具体的に提案して相手を説得してください。その後も会話を続けます。	

注：どちらのタスクでも、「～分目安。以下の設定で会話をしてください。」との指示があり、以下のように、名前とタスク名を言ってから始めるように指示している。これがないと録音を聞きながらの評価が困難になる。

Student A: My name is (　　). I am Student A.

Student B: My name is (　　). I am Student B. This is Conversation 7, Toothache.

ロールプレイでは、誰が話し始めるのかの指定があり、ディスカッションでは、「どちらからでもよいので、話し始める」ように指示してある。

Toothache は平均的な難易度、Hobby は容易という結果だった。

③研究 3 (学会発表③)

プレテストとポストテストの間に、3 回形成的テストを数回受けた実験群（51 名）と、プレテストとポストテストのみを受けた統制群（30 名）のクラスを比較することにより、ロールプレイテストの再テスト法による信頼性と、形成的テストを定期的に受けることによる大学生への心理面への波及効果を調べた。

④研究 4 (学会発表②)

日本人大学生においてロールプレイタスクとディスカッションタスク（計 11 タスク）で頻繁に使われる発話機能の種類を調べた。会話で実際に使われている発話機能と使われていない発話機能は何か、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、タスクとトピックを系統的に設定するのに有用な知見を得ようとした。各タスク 20 名分の発話を書き起こし、使用されている発話機能のコーディングを行った。

⑤研究 5 (学会発表①)

研究 2 で作成した総合的尺度に加えて「分析的尺度」を作成し、その妥当性を、多相ラッシュ分析とピアソン積率相関係数を用いて検証した。分析的尺度は総合的尺度と同様に 3 段階とし、その観点は、発音・イントネーション、文法と語彙、流暢さ、やり取りによるコミュニケーション、タスク到達度の 5 点とした。

110 名の大学生が計 10 個のタスクに取り組み、3〜4 名の評価者が、同じ発話を総合的尺度と分析的尺度の両方で、別々に評価した。

⑤研究 6 (学会発表⑤で一部発表; 枠組みを図書①として発行)

研究 1〜5 で得た妥当性の証拠を、Chapelle et al. (2008)の枠組みに沿って整理し、包括的な妥当性検証のために今後必要な点をまとめた。

４．研究成果

①研究 1

タスクと評価者は予測されるモデルに適合しており、ルーブリックも適切に機能していた。1 つのタスクを実施して 2 名の評価者が採点する場合か、3 つのタスクを実施して 1 名の評価者が採点する場合に、高い信頼性が得られた。ロールプレイとディスカッションのタスクは似た能力を測っていて、全体として一次元構造が見られたことなどが示された。

②研究 2

タスクと評価者は予測されるモデルに適合していた。ルーブリックも適切に機能していた。ロールプレイとディスカッションのタスクは、2 つは似た能力を測っており、一次元構造があることが示された。タスクの難易度の幅は限定的で、より難易度が高いもの、低いもの、中間レベルに位置するものをさらに追加する必要性が指摘された。十分な信頼性を保つためには、タスクを 4 つ行って 2 名で評価することが必要であることも示された。ロールプレイの場合、表 1 にあるようにカードが人によって異なるが、そのカードによって難易度が異ならないかを調べたところ、ほとんど変わらなかった。

③研究 3

テスト得点ではプレテスト・ポストテスト間でほぼ違いがなく、再テスト法による信頼性の高さが示された。心理面では、統制群では英語学習への動機づけが下がったのに対し、実験群では動機づけが保たれていた。これにより、形成的テストを授業で受けることにより、動機づけ減退を妨げる効果があることが示唆された。

④研究 4

分析の結果、全体的に発話機能の出現は限定的であることが分かった。発話機能とタスクタイプの関連について、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ていた。ディスカッションでは「相手が言ったことに対して応答する」機能がより多く引き出されていた。これらは、ロールプレイやディスカッションのタスクで引き出しやすい機能だと考えられる。逆に、引き出す意図で作成したが引き出されなかった機能もあった。その例は、旅行に持参するものを決めるタスクにおいて、「異議を唱える」「詳細に述べる」「個人的な情報を述べる」機能であり、夕食に誘われるが辞退して会話を続けるタスクにおいて、「意見や行動の理由を述べる」機能であった。今後、熟達度が高い受験者においても同様かを調べ、意図した機能が引き出されていな

ければ、その点での妥当性が低いことになり、今後のタスク修正等が必要になるだろう。

⑤研究 5

総合的尺度と分析的尺度はともに、ほぼ適切に機能していた。改善点は、総合的尺度のレベル 1 とレベル 3 の違いが大きすぎたこと、分析的尺度の 2 観点（やり取りによるコミュニケーション、タスク達成度）において、分析モデルが予測するパターンと異なるものが見られたことであった。

分析的尺度の 5 観点を合わせた結果と、総合的尺度の結果の相関を調べたところ、.85 と非常に強く、かなり共通した側面が測れることが示された。

分析的尺度の 5 観点の数は多く、実践で使用する際には負担になる可能性がある。そのため、観点を減らしていき、総合的尺度の結果の相関をそれぞれ調べていったところ、表 2 のような結果になった。なおこの分析の際には、ロールプレイやディスカッションのタスクで測る能力として重要な「やり取りによるコミュニケーション」と「タスク到達度」の 2 つは入れることを前提に、その 2 観点と他の観点との組み合わせを調べた。

表 2．総合的尺度と、観点を変えた分析的尺度の相関

	総合的尺度	分析的尺度							
		5 観点	4 観点			3 観点			2 観点
		PGFIT	GFIT	PFIT	PGIT	FIT	GIT	PIT	IT
総合的尺度	--	.85	.86	.85	.85	.87	.86	.84	.85
受験者の信頼性	.92	.96	.95	.94	.95	.93	.94	.94	.88
タスクの信頼性	.92	.96	.96	.94	.93	.94	.94	.89	.90
尺度の信頼性	--	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99

注：P = 発音・イントネーション．G = 文法と語彙．F = 流暢さ．I = やり取りによるコミュニケーション．T = タスク到達度．多相ラッシュ分析を用いたため、複数の信頼性が算出されている。

結果として、相関はすべて .84 以上であり、5 観点から 2 観点まで減らしても、総合的尺度との関係という意味ではほとんど変わらないことが示された。また総合的・分析的尺度における信頼性を調べたところ、どの場合でも .88 以上と非常に高い信頼性が示された（表 2 参照)。

そのため、分析的尺度の観点は 2 つまで減らしても信頼性が十分保たれ、総合的尺度と似た側面を測れる程度がほぼ同じため、観点の削減が可能であり、テストが測る構成概念から考え、「やり取りによるコミュニケーション」と「タスク到達度」の 2 観点を入れた分析的尺度を使うことが提案された。

⑥研究 6

Chapelle et al. (2008) の論証に基づく妥当性主張の枠組みにおいては、領域定義、得点化、一般化、説明、外挿、利用の 6 段階の推論がある。本研究では、以下の証拠を得た。

(1) 領域定義：専門家判断によって確認された、タスクの関連性と代表性（プラスの証拠)
(2) 得点化：多相ラッシュ分析によって確認された、タスクとルーブリックの統計的特徴（プラスの証拠）、幅広い難易度のタスクが十分にはなかった点（マイナスの証拠）
(3) 一般化：多相ラッシュ分析によって確認された、テストや評価者の高い信頼性（プラスの証拠)
(4) 説明：確認的因子分析によって、意図したように一次元の能力を測る構造が確認された点(プラスの証拠)、発話機能分析によって示された、テストが引き出す機能が限定的であった点（マイナスの証拠)
(5) 外挿：相関分析によって確認された、全体的な英語力を測るテストと予想通りの程度の相関（プラスの証拠)
(6) 利用：実験的分析によって確認された、テストの実施による学習者の動機づけの安定性（プラスの証拠)

今後調べる予定の観点は以下である。

(2) 得点化：幅広い難易度のタスクを多く作り、タスクの難易度の幅を意図通り広くできるか
(3) 一般化：学生同士が話すときのどんな人と組むかによるテスト得点への影響がどの程度あるか
(4) 説明：英語熟達度が高い受験者でもテストが引き出す機能が限定的であるか
(6) 利用：テストを行うことでの学生の英語学習やスピーキングに対する態度を、多様な観点で長期間調べたときに、どのような影響が見られるか

最終的に、得た様々な観点からの証拠に基づき、論証に基づく妥当性主張を行い、ロールプレイテストの特徴として提示していく。このテスト方法の普及については、既に図書②、講演・研修等で行っており、さらに活動を続けていく。

５．主な発表論文等

〔雑誌論文〕（計 41 件）

① Koizumi, R., Okabe, Y., & Kashimada, Y. (2017). A multifaceted Rasch analysis of rater reliability of the Speaking Section of the GTEC CBT. *Annual Review of English Language Education in Japan*, *28*, 241-256. doi:10.20581/arele.28.0_241 査読有

② Koizumi, R., In'nami, Y., Asano, K., & Agawa, T. (2016). Validity evidence of Criterion® for assessing L2 writing proficiency in a Japanese university context. *Language Testing in Asia*, *6*(5), 1-26. doi:10.1186/s40468-016-0027-7 査読有

③ Koizumi, R., & In'nami, Y. (2014). Modeling complexity, accuracy, and fluency of Japanese learners of English: A structural equation modeling approach. *JALT Journal*, *36*, 25-46. Retrieved from http://jalt-publications.org/jj/articles/3728-modeling-complexity-accuracy-and-fluency-japanese-learners-english-structural-equat 査読有

〔学会発表〕（計 56 件）

① Koizumi, R., In'nami, Y., & Fukazawa, M. (2019, March). *Holistic and analytic scales of a paired oral test for Japanese university students*. Poster presented at the 41st Language Testing Research Colloquium, Courtyard Atlanta Decatur Downtown/Emory, Atlanta, USA.

② 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. (国際学会)

③ Koizumi, R., In'nami, Y., & Fukazawa, M. (2016, September). *Effects of four-month paired-oral-type instruction and assessment on the development of L2 speaking ability of Japanese university learners of English*. In K. Saito (Chair), The longitudinal development of L2 ability in Japanese EFL classrooms. Symposium conducted at the Pacific Second Language Research Forum 2016 (PacSLRF2016), Chuo University, Tokyo, Japan. (国際学会)

④ Koizumi, R. (2016, August). *Multi-faceted Rasch analysis in rating tasks in Japan's university entrance examinations*. Invited presentation at Pacific Rim Objective Measurement Symposium (PROMS) 2016. Orient Hotel, Xi'an, China. (国際学会、招待講演)

⑤ Koizumi, R., In'nami, Y., & Fukazawa, M. (2015, September). *An argument-based validation framework of a paired oral test in university classroom contexts*. Paper presented at the 19th Annual Conference of the Japan Language Testing Association (JLTA), Chuo University, Tokyo, Japan.

〔図書〕（計 12 件）

① 小泉利恵（2018）．『英語 4 技能テストの選び方と使い方―妥当性の観点から―』 アルク (総 263 ページ)

② 小泉利恵・印南洋・深澤真（編著）．(2017)．『実例でわかる英語テスト作成ガイド』 大修館書店（総 161 ページ）

③ 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, 総 424 ページ). Gateway East, Singapore: Springer Singapore. doi:10.1007/978-981-10-1687-5

④ Koizumi, R., In'nami, Y., & Fukazawa, M. (2016). Development of a paired oral test for Japanese university students. In C. Saida, Y. Hoshino, & J. Dunlea (Eds.), *British Council New Directions in Language Assessment: JASELE Journal Special Edition* (pp. 103-121, 総 166 ページ). Tokyo: British Council Japan.

〔その他〕

ホームページ等

http://www7b.biglobe.ne.jp/~koizumi/KoizumiHP.html

６．研究組織

(1)研究分担者
研究分担者氏名：深澤　真
ローマ字氏名：(FUKAZAWA, Makoto)
所属研究機関名：琉球大学
部局名：教育学部
職名：准教授
研究者番号（8 桁）：00634429

研究分担者氏名：印南　洋
ローマ字氏名：(IN’NAMI, Yo)
所属研究機関名：中央大学
部局名：理工学部
職名：教授
研究者番号（8 桁）：80508747