様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月　５日現在

機関番号：１３９０１
研究種目：基盤研究(C)
研究期間：2010　～　2012
課題番号：22520528
研究課題名（和文）
大規模試験としての日本語口頭能力測定における評価システムに関する研究
研究課題名（英文）
Developing a web-based rating system for Japanese Oral Proficiency Test as large scale test.
研究代表者
野口　裕之（NOGUCHI HIROYUKI）
名古屋大学・教育発達科学研究科・教授
研究者番号：60114815

研究成果の概要（和文）：日本語口頭能力試験のための評価システムを開発する研究を進めた。このシステムではWebベースで配信した受験者の発話標本を聞きながら、評価者がPC画面上に逐次提示される評価票の各項目に評定結果を入力する。評価者毎には課題を通して比較的一貫した評定であった。評価者間では「量的評価」の方が「質的評価」よりも相対的に一致した結果が得られたが、「量的評価」でも一部の評価者で他と異なる評定結果を示した。多相ラッシュ分析を適用した結果は、評価者の厳しさの違いは無視できないが、推定された能力尺度値は予め12名の発話者に想定した能力水準と大きくは異ならないことを示した。

研究成果の概要（英文）：The present study developed the rating system for Japanese Language Oral Proficiency Test. In the system, spoken samples of examinees' speech were accessed through the Internet. Raters could enter the results for each item shown on a scoring table, which was successively indicated on a PC screen. Rating of each rater was relatively consistent for each domain. Quantitative ratings were relatively more consistent among raters than qualitative ratings. However, some raters gave different quantitative ratings from others. The results of multi-faceted Rasch analysis indicated that dispersion of strictness between raters was significant. The estimates on the ability scale were not significantly different from the assumed ability of the 12 speakers.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010年度	1,500,000	450,000	1,950,000
2011年度	900,000	270,000	1,170,000
2012年度	600,000	180,000	780,000
総　計	3,000,000	900,000	3,900,000

研究分野：人文学
科研費の分科・細目：言語学・日本語教育
キーワード：言語テスト、口頭能力測定、評価システム、評価票、多相ラッシュ・モデル

１．研究開始当初の背景

本研究では、大規模試験の一部門となることを想定して、PC による日本語口頭能力試験の開発に関わる諸問題を取り上げている。元々は、日本語能力試験企画小委員会口頭能力試験調査部会(2003)で開発されたテストをベースとしている。このテストは、対面式ではなく、ＰＣでテスト課題を提示し、受験者がそれに応答するという方式であり、これまでに、課題の提示方法の妥当性(庄司・青山・金澤・伊東・野口、2003)、分析的な評価方

法の妥当性(庄司・野口・金澤・青山・伊東・迫田・春原・廣利・和田、2004)の基礎研究の結果が報告されて来た。ただし、これらの研究で扱われた口頭能力には、他者との交互作用(interaction)は含まれず、基礎的な発話能力を対象としている。CEFR では外国語能力のうち Speaking を Spoken Interaction と Spoken Production に下位区分しているが、これらの基礎研究で取り上げられた能力は,後者の Spoken Production が該当する。

日本語能力試験の開発主体である国際交流基金の関与は 2003 年までで、その後は日本語能力試験企画小委員会口頭能力試験調査部会の一部の関係者に加えて、このような試験の開発に関心を持つ専門家から構成される研究チームが実用水準を念頭において開発を継続して来た。

一方、最近の PC の性能の向上や利用技術の発展には目覚ましいものがあり、日本語教育の諸領域でもコミュニケーション媒体や学習のための道具として急速に普及し，教育現場や個人の学習環境において誰もが簡単に入手し、利用できる便利な道具となっている。

このため、日本語口頭能力測定に PC を利用する方法に関しても、以前に比べて可能性が大きく拡がっている。例えば、採点のために PC が多数用意されたセンターは必ずしも必須ではなく、情報のセキュリティが確保できるならば、評価者個人の作業場所にテストの結果を配信して、評価者は評価結果を送り返すということも可能な状況になっている。それにも関わらず、一定規模の受験者を対象とした日本語口頭能力試験は実用水準に至ってはいなかった。

２．研究の目的

受験者の発話を評価者に WEB ベースで配信して、評価者が発話を聞きながら、PC 画面上に表示される評価票に評定項目に対する評定結果を入力し、最終結果を返信するという、「日本語口頭能力評価システム」を構築することを最終目的とした。

そのために、

(1)日本語口頭能力試験の発話課題を開発し、確定すること、
(2)発話者の口頭能力を評価するための評価票を開発（洗練）すること、
(3) PC による日本語口頭能力試験に対する受験者の評価を得て、試験の当事者の視点から何に注目しどのように試験の改良につなげていくべきなのかを検討すること、
(4)WEB ベースで受験者の発話を評価者に送信して、評価者が評価票に基づいて評定した結果を入力し、それらの結果を集積して、分析するシステムを開発すること、
(5)評価者の厳しさの散らばりがどの程度あり、それを考慮したモデルによる採点の可能性を検討すること、

という、下位目的に分けて研究を実施した。

３．研究の方法

研究の目的の下位目的に対応させて、方法を述べる。

(1) 日本語口頭能力試験発話課題の検討

① 日本語口頭能力試験の発話課題には、先行研究では、「Q&A(質問に短く答える課題)」、「留守番電話(メッセージを吹き込む課題)」、「絵説明(4 枚の絵のストーリーを体験談として話す課題)」、「議論を聞いて意見を述べる(あるテーマについて自分の意見を述べる課題)」の４課題であったが、今回、受験者が自ら質問をして必要な情報を収集する能力を測定する「情報収集(わからないことについて質問する課題)」を加えたため、日本語学習の中級者および上級者から協力者を募り、PC から音声と画像とで提示される課題に対して実際に回答（発話）してもらい、それを録音する。
②仮評定票を用いて、量的側面、質的側面の両面から各発話者について評定する。この段階での評定は研究チームの日本語教育専門家が実施する。
③これらの結果から、課題の性能について検討する。
④５つの課題の測定する構成概念について、Fulcher(2003)、Bachman & Palmer(1996)、清水(2009)などを参照して、理論的枠組みから検討する。

(2)評価票の検討

①評価票は「量の判定」と「質の判定」から構成される。「量の判定」では、受験者がたくさん話したかどうかではなく、構成概念から抽出された言及事項が一致するかどうかを評価するものである。ここでは発音はどうか、文法はどうかなどの発話の質については評価しない。また、「質の判定」では、即応性と滑らかさ、発音のわかりやすさ、語彙のわかりやすさ、文の構造のわかりやすさなどについて、発話を聴いて評価する。これらの各項目に対して、課題達成の度合いを 3 段階から 4 段階に分けて記述し、評定尺度を構成した。評価者はこの評定尺度を参照しながら項目ごとに発話者の課題達成度を決める。
②発話者は(1)と同一の協力者である。
③評価者は研究メンバーの中の日本語教育専門家 4 名である。
④4 名が評定した結果を統計処理し、2 名の間の評価者間の一致度を検討する。

(3)受験者の視点からの検討
①受験者からの評価に焦点を当て，それを分析し、テストをどのように捉えたのかを考察することにより、受験者側の視点からの試験の改良点を検討する。
②発話者は(1)と同一の協力者である。
③属性は、大学または大学院の留学生で，母語の内訳は，中国語 4 名，韓国語 2 名，ベトナム語 2 名，英語，ジャワ語，モンゴル語，蘭州語がそれぞれ 1 名ずつである。また，日本語能力のレベルは中上級から上級で，ほとんどが日本語能力試験の N1(旧 1 級)あるいは N2 レベルの合格者である。
④(1)のテストに続いて、1 対 1 の対面式でインタビュー(平均 11 分)を行った。インタビューは，予め用意した項目に基づき半構造化形式で進め，そのやりとりを IC レコーダーで録音する。このインタビューの文字化資料に対して QDA ソフトを用いて定性的な分析を行なう。
(4)システムの開発
①評価システムの開発は、主として研究メンバーの中で言語情報科学の専門家が実施した。
(5)評価者の厳しさの違いの検討
①評価協力者として応募下さった 53 名の方々に、評価者 1 名あたりに発話者 3 名の発話標本の評定を依頼した。発話者の割り当てにあたって、日本語の能力水準が偏らないように配慮した。なお、評価協力者の応募資格として、
・インターネットに接続できる PC を使える方
・日本語を第一言語とし、かつ、日本国内に在住していらっしゃる方
・以上の条件を満たし、次のいずれかに該当する方
　ⅰ）日本語教育歴 3 年以上ある方
　ⅱ）ACTFL-OPI テスター資格（日本語）をお持ちの方
ということを条件とした。なお、ACTFL-OPI テスターである、という条件はスピーキング・テストの経験があることを保証する以上の意味はここでは持っていない。
②評価協力者には、「ご自宅などの PC で Web 上の日本語口頭能力評価システムにアクセスし、所定の評価方法に従って日本語学習者 4 名（練習 1 名、本評定 3 名）の発話を評価していただきます。作業工程の一部でハードコピーの評価表を使用します。」ということを予め伝えた上で応募いただいた。
③評価協力者による評定結果は、評価者 ID および発話者 ID のあとに「量の評価」「質の評価」各項目に対する評定結果が 1 行に並ぶようなデータ行列に整理した。
④発話者別に複数の評価者の結果を統計分析する。
⑤評価者要因をモデルに組み込んだ、多相ラッシュ・モデルにより評価者の厳しさの違い、その違いを調整した場合の発話者の能力値について検討する。

４．研究成果

研究の方法に対応させて研究成果を述べる。

(1) 日本語口頭能力試験発話課題の検討
[発話課題の検討]
①予備実験の結果、「絵説明」課題は，個人の経験談としたためか発話が得やすくなったが，情景描写・心理描写は難易度が高く，能力差が現れた。
②新しく導入した「情報収集」課題は，中級者・上級者とも一定の発話量が採取できたが，受験者間に質的な差が出にくいことが明らかになった。これは、発話には質問を羅列しただけの不自然なものが多く，評価の優劣がつけ難いということによるものであった。その原因は，課題遂行においてⅰ）自分の役割が十分に認識できていない，ⅱ）質問をする対話者の存在が想定しにくい，という点にあるのではないかと考え，課題の教示方法について，ⅰ）音声と文字の両方を用いて役割を明示する，ⅱ）対話相手の顔を表示する，という改良を加えた。その結果，本調査では相手へ働きかける発話や、接続表現を用いたまとまりのある発話がみられ，対話者を想定したと考えられる発話が増えた。
[構成概念の検討]
③発話課題の構成概念について
構成概念の定義のプロセスは 4 つのステップに分けて実施した。全てのプロセスは、5 名の研究メンバーで定期的に会議およびメールでの意見交換を行いながら進められた。
ⅰ）理論的枠組みの検討：理論的枠組みとして、スピーキング能力の構成要素に、言語能力、方略的能力、テクスト的知識、語用論的知識、社会言語的知識の 5 つの下位能力を設定する Fulcher(2003)を採用し、その枠組みをもとに検討を行った。ただし、Fulcher (2003)では社会言語的知識についての説明が少なく、語用論的知識との区別が曖昧であったため、Bachman & Palmer(1996)や清水(2009)を参照し、語用論的知識の下位要素として社会言語的知識が含まれるようにした。また、語用論的知識には機能的知識も含まれるが、その機能の分類として尾崎・椿・中井(2010)の分類を参照した。
ⅱ）測定対象項目となるかどうかの検討：このように整理した理論的枠組みを用いて、次に、各構成要素が本テストで測定対象となるかどうかを検討した結果、方略的能力は測定対象とはならないとした。受験者は、解答

の際に様々な方略的能力を使用すると考えられるが、本テストで得られる発話データのみから方略的能力を使ったかどうかの判断は難しいと考えたためである。また、テクスト的知識と語用論的能力は、テスト課題の種類により測定対象となる場合とならない場合があることがわかった。例えば、Q&A の課題では質問に対する比較的短い応答となるため、テクスト的知識を測定するのは難しいと考えられる。

ⅲ）新たに追加する観点の検討： Fulcher (2003)では拾いきれなかった、他の観点がないかどうか検討を行った。研究メンバー内での意見交換や、過去の先行調査時に実施した評定者アンケートの見直しを行い、追加する観点の候補が挙げられたが、それらは何らかの形で Fulcher(2003)の枠組みに含まれていると判断し、最終的には新たな観点の追加は行わなかった。

ⅳ）テスト課題ごとの測定対象要素：本テストは複数の課題で構成されているため、テスト課題ごとに測定対象要素を整理した。すなわち、a)言語能力は５つの課題の全てで測定するが、b)方略的能力は全ての課題で測定しない、c)テクスト的知識は Q&A では測定しないが、その他の４課題では測定する、d)語用論的知識は５つの課題全てで測定するが、その下位能力の機能的知識については課題によって測定対象とする要素が異なっている。

以上の検討の結果、各課題がそれぞれ異なる機能を測定しており、テスト全体として概ねバランスが取れていることがわかった。

(2)評価票の検討

①「Ｑ＆Ａ」、「留守番電話」、「絵説明」、「議論」、「情報収集」の５つの課題毎に評価者間の評定について相関係数を計算した。評定の一致度を見る場合に相関係数を用いることが適当ではない場合もあるが、本研究では相関を見る評価者間で評定の平均値が大きく異ならないため、変量間の関連の度合いをみるのによく用いられる相関係数を用いた。

質的評価ではおおむね高い数値が得られたが、量的評価では「絵説明」課題について相関係数が 0.467 であり、他の課題に比べて低い値が得られた。その原因は、研究チームが「絵説明」をストーリーテリングととらえていたのが原因ではないかと考えた。本来この課題で期待した発話にはストーリーテリングに経験談を含めたものであった。ストーリーテリングは起きた出来事を忠実に描写する必要があるのに対し、経験談は話し手が最も前景化したい部分に時間を使う。この出来事で一番前景化したいのは話の展開が起こる 2 枚目と 3 枚目が中心だと思われる。受験者の発話そのものでも、2 枚目と 3 枚目を詳しく話している人が多い。このことから前景化する部分での発話を十分に捉えられるような評価票を作成するということである。

②以上のことから、評価票の改善を行った。具体的な改善箇所としては、まず上位カテゴリーを作成した。改善前は「いつ」「どこで」「誰が」などの言及事項を並べただけであったが、改善後は事件を軸にした展開をカテゴリー化した。カテゴリー化したことにより視覚的にも見やすくなり、発話に評価がついて行けないということがなくなった。

③次に、受験者の発話をもとに項目を見直し、その結果、項目を追加したり、項目をより細分化した。改善前の項目は「どうした」というものが 1 つしかなかったが、「猫がコップをひっくりかえした」と「コーヒーがこぼれた」を分けた。

④改善後の評価尺度で「絵説明」の量を再度評価した結果､相関係数が 0.593 に上昇した。評価項目を細分化、精査したことで、評価すべき点が明確になり、評価者間のずれが小さくなった。

⑤以上の結果から、改良された評価尺度は概ね機能すると考えられる。

(3) 受験者の視点からの検討

分析の結果，上位の概念カテゴリーとして〈課題の達成感〉，〈試験に対する肯定的情意〉，〈試験に対する否定的情意〉，〈試験環境への適応〉，〈今後への期待〉等が生成された。〈課題の達成感〉では，「難しい」，「上手く出来なかった」とされた課題について，その要因として，それを説明する言葉が出てこない，語彙が足りないといった自分の日本語能力について語る場合と，その課題の特性を実際の場面と比較して語る場合，さらに音質などのインターフェースについて語る場合がみられた。〈課題の達成感〉が，実際は違う，普段考えたことがないので答えにくいといった〈試験に対する否定的情意〉に結びつく場合は，自分の能力への言及はあまりみられなかった。一方，上手く出来なかったとしながらも，日常よくある場面で答えやすい，自由に話せて楽しい，面白いなどの〈試験に対する肯定的情意〉に結びつく場合では，自分の日本語能力の不足部分についての言及がみられた。また，PC に向かって話すという環境については，違和感，不自然さがあるとする者もいたが，現行の日本語能力試験には口頭能力試験がないこと，そのために口頭能力には自信が持てないとし，多様な課題を採用している本試験システムへの期待の高さが窺えた。

(4)システムの開発

評価システムは、あらかじめ接続先と ID が知らされている評価協力者が WEB 上で評価システムにアクセスして、評価の説明、練習

（発話者１名分)、本評定（発話者３名分）と進むように設計された。

評価システムは、試行を繰り返しながらトラブルの原因を取り除いたり、画面の見やすさなどについても検討を進めながら完成させた。評価協力者の評価作業中に大きなトラブルが発生することはなかった。

(5)評価者の厳しさの違いの検討

①発話者別に複数の評価者の結果を統計分析する。

全部で 12 名の発話者に対して。それぞれ 12 名ないし 15 名の評価者が評定した結果を、各評定者毎に量的評価、質的評価についてそれぞれ合計した値を分析した。

各発話者の基本統計量は表 1-1、表 1-2 に示した通りである。また、量的評価、質的評価の各合計値の分布状況を図 2-1、図 2-2 に箱ヒゲ図で示した。

平均値で見ると、当初研究グループで想定した発話者の能力水準とほとんど変わらない結果が量的評価、質的評価の両方で得られた。

また、全体として評価者による評定結果の違いは一定程度存在し、量的評価の方が質的評価より標準偏差が小さく、評価者間の一致度が高いことが示された。一般に評定者間の一致度の高い量的評価において、発話者番号 10 の標準偏差が 4.8 と他と比べて極端に大きな値を示しているが、この発話者の質的評価に関しては、特に他の発話者と標準偏差が異なる値は示していない。

表１－１　量合計　記述統計量

発話者番号	度数	最小値	最大値	平均値	標準偏差
1	12	28	32	29.6	1.5
2	12	29	34	31.0	1.6
3	13	27	34	30.2	1.9
4	13	25	32	28.2	2.1
6	13	28	32	30.2	1.0
7	12	30	37	34.8	2.0
8	13	24	29	26.2	1.8
9	13	27	33	29.5	1.9
10	13	16	34	23.8	4.8
11	13	22	27	24.9	1.7
12	14	27	34	30.8	1.9
13	15	31	35	33.1	1.3

表１－２　質合計　記述統計量

発話者番号	度数	最小値	最大値	平均値	標準偏差
1	12	74	118	96.3	14.6
2	12	53	116	83.3	20.4
3	13	49	104	72.5	14.7
4	13	45	111	80.3	16.8
6	13	63	102	80.3	11.1
7	12	84	125	111.2	12.2
8	13	23	83	58.1	15.9
9	13	61	123	93.7	14.6
10	13	41	103	66.2	18.5
11	13	37	89	65.0	16.1
12	14	78	122	100.6	14.4
13	15	87	127	110.9	10.5

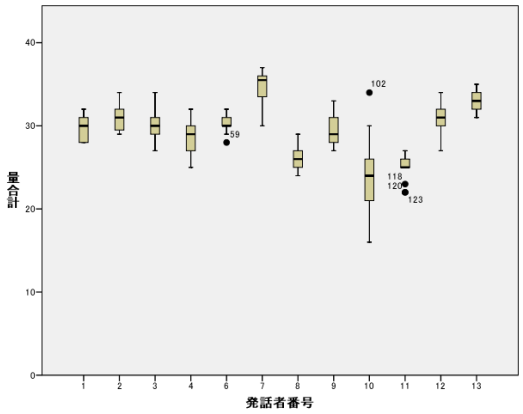

図 2-1　量的評価合計値の分布状況

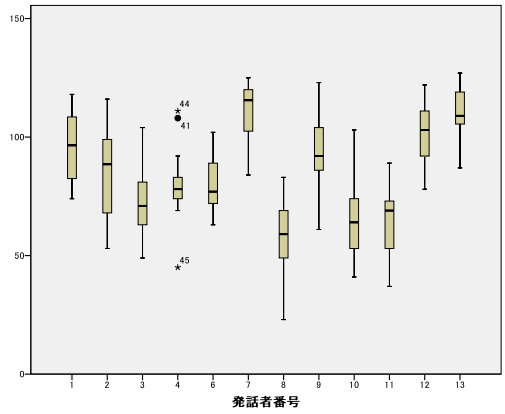

図 2-2　質的評価合計値の分布状況

発話者番号 10 の発話標本について詳細に検討したところ、発音の正確さに問題があり、評価者により発音の正確さを重視する程度が異なっていることが反映したものと推測される。評価者がどのような日本語教育機関に属するか、日頃どのような母語話者に教えることが多いか、などが評定に影響することが明らかになり、今後評定者に対する研修で

配慮すべき課題である。

②評価者要因をモデルに組み込んだ、多相ラッシュ・モデルにより評価者の厳しさの違い、その違いを調整した場合の発話者の能力値について、Bond & Fox Facets を利用して分析した結果、評価者要因は無視できない程度に厳しさの違いが認められるが、発話者の能力尺度値に関しては、当初研究グループで想定した能力水準と順序性に大きな違いは見られなかった。

(6)得られた成果の位置づけと今後の展望

日本語の口頭能力を測定する大規模公的試験が存在しない中で、本研究の成果は日本語能力試験のような大規模公的試験の中に取り入れて実施することが期待される。

本研究で採用した評価方式は、音声認識による自動採点システムとは異なり、多くの受験者データを一定期間内に処理して、結果を返却することはできない。そのため、受験者数に何らかの制限が必要になる、その点が難点とも言えるが、しかしながら、研修を受けた評価者が、量的側面と質的側面の両側面から丁寧に評定して積み上げた結果であるため、きめの細かい評価結果になっている。英語の大規模能力試験で有名な TOEFL はコンピュータで、IELTS は研修を受けた面接官がスピーキング能力を測定している。当該試験が意図する測定目的に応じて方法が異なって来る。今後、本研究はさらに発話標本を増やして実験を続けることと、課題の枠組みや構成概念はそのままで、具体的な課題を取り替えて試験を実施するための方策を探ることが次の課題である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 6 件）

①野原ゆかり・小林久美子、PC による口頭能力試験に対する受験者の評価－受験当事者は試験をどのように捉えたか、日本語教育学会 2013 年度春季大会、2013 年 5 月 26 日（発表予定、採択済）、立教大学

②安高紀子・堀川有美、パソコンによる口頭能力試験における新課題の改良、科学研究費助成事業合同成果発表会「言語能力評価の最前線～運用力の評価を目指して～」、2013 年 3 月 28 日、桜美林大学

③小林久美子・安高紀子・野原ゆかり、パソコンによる日本語口頭能力テストの評価尺度、第 45 回　お茶の水女子大学　日本言語文化学研究会、2012 年 12 月 8 日、お茶の水女子大学

④堀川有美・小林久美子、パソコンによる日本語口頭能力テストの構成概念の定義、第 45 回　お茶の水女子大学　日本言語文化学研究会、2012 年 12 月 8 日、お茶の水女子大学

⑤庄司惠雄・安高紀子・和田晃子・野口裕之、ＰＣによる口頭能力試験のための新課題開発および課題改良、日本語教育国際大会、2012 年 8 月 18 日、名古屋大学

⑥和田晃子・堀川有美・小林久美子・野口裕之、コンピュータ・ベースの日本語発話能力評価システム-WEB 方式による評価の試み、日本語教育学会 2010 年度秋季大会、2010 年 10 月 10 日、神戸大学

６．研究組織

(1)研究代表者

野口　裕之（NOGUCHI HIROYUKI）
名古屋大学・大学院教育発達科学研究科・教授
研究者番号：60114815

(2)研究分担者

堀川　有美（HORIKAWA YUMI）
桜美林大学・大学院言語教育研究科・客員講師
研究者番号：50557171
(H22,23→H24：連携研究者)

李　在鎬　(RI JEHO)
桜美林大学・大学院言語教育研究科・客員講師
研究者番号：20450695
(H23→H24：連携研究者)

(3)連携研究者

庄司　惠雄（SHOJI YOSHIO）
前お茶の水女子大学・留学生センター・教授
研究者番号：40253017

熊谷　龍一（KUMAGAI RYUICHI）
東北大学・大学院教育学研究科・准教授
研究者番号：60422622

野原　ゆかり（NOHARA YUKARI）
国立国語研究所・プロジェクト奨励研究員
研究者番号：30584578