口頭テストにおける観点別評価
—日本語教師による評価と発話との関係について—

仁科　浩美

キーワード：口頭テスト、評価項目、相関、重み付け

要　旨

「話す」技能を養うことは研究留学生にとっての大きな課題である。そのため、教師が適切に評価を行い、学習者に習得状況について情報を与えることは重要である。本稿では、作成した口頭テストの結果をもとに、発話データと教師の評価とを比較し、教師が学習者の発話をどのようにとらえているのかを７項目について分析した。その結果、「沈黙」や「文の型」など量的にとらえやすい項目では相関が見られたが、評価と発話データとは、必ずしも一致するわけではなく、「文法」「発音」では質的要素が関与し、１つの誤りの中に重み付けがあることが分析により裏付けられた。

１．はじめに

日本の大学院レベルで研究を行う留学生、すなわち、研究留学生にとって「話す」技能の必要性は高い。そのため、日本語学習の課程において、日本語教師が口頭能力に対し形成的評価を与えることは非常に重要である。それは、ニーズが高いという理由と同時に、通常、一般の日本人が非母語話者の発話に対し、直接的に助言したり、評価したりすることはほとんどないということからも言えることである。

しかし、教師の評価は学習者の発話をどのようにとらえているのだろうか。本稿では、口頭テストにおける実際の発話と教師評価との関係を観点別に分析する。

２．先行研究

教師を対象に口頭テストの評価と発話の実際とを比較した研究は多くない。Douglas（1994）は、英語教師に発音・文法・流暢さ・わかりやすさについて評価させ、評価平均値が同じスコアでも、発話は質的に異なることを具体例に基づき説明した。しかし、評価者が２名と少なく、評価傾向が十分には把握できない。

また、鈴木（2004）は、発話評価と日本語教師経験の長さとの関係を分析した。その結果、観点別では中級テープの評価において「文法」に経験の多少と評価に差がみられたが、上級テープ評価では項目すべてに差がみられなかったと報告しており、教師経験が評価に与える影響はさほど大きくないことがうかがえる。

石崎（1999）は、教師ではなく、日本語母語話者（20代女子大学生）を対象に母語話者による「主体的評価」と、発話を一定の基準で数量化した「客観的評価」との関係を分析した。その結果、「語彙」と「音声´(不自然な音の高低）」に関しては、有意に相関が見られたが、「語・文法の誤り」「音声（音節の誤り）」については、相関は有意ではなかったと述べている。有意でなかった理由の一つとして、非教師が評価者であることをあげている。

本稿は、これらを踏まえ、先行研究でまだ扱われていない、日本語教師（12名）の評価結果を発話と比較し、その関係を検討するものである。

3．口頭テストとその評価の概要

3．1．口頭テスト概要

口頭テストは、平成17年6月と平成17年7月に2回行った。前者をテスト1、後者をテスト2と呼ぶ。テスト受験者は東北大学国際交流センターに在籍していた初級学習者8名である。評価者は同等レベルの学習者への指導経験を持つ母語話者日本語教師12名である。テストの所要時間は、10分程度を目安に行ったが、タスクを達成することを重視したため、特に厳しい時間制限は設けなかった。平均所要時間はテスト1で7分47秒、テスト2で9分07秒である。

テストの内容(注1)は、テスト1・2ともそれぞれ3つのタスクから構成される。

①タスク1：試験官からの質問に答える。

②タスク2：ある話題について、受験者が一方向的に説明する。

③タスク3：試験官とのロールプレイ

3．2．評価項目と評価方法

テストの評価項目は、タスクが達成されたかどうかを見る包括的評価と、個々の観点がどうであったかを見る観点別評価とに分かれるが、本稿では、後者についてのみ検討する。評価項目は、過去に公開されている日本語教育機関（加納2004；庄司1996；因1994等）での口頭テストを参考に、正確さ・流暢さ・不快さの3つの観点から、8つの評価項目を設けた（表1）。

表１　評価項目

観点	評価項目
正確さ	a. 語彙　：必要な語彙を間違うことなく適切に使うことができるか。 b. 文法　：間違うことなく使用できるか。 c. 発音　：音・イントネーションに気になる不自然なところがないか。
流暢さ	d. 沈黙　：沈黙が頻繁に見られるか。 e. 反復　：反復（繰り返し）が頻繁に見られるか。 f. 発話量：この段階として発話量は十分か。 g. 文の型：どの程度の構文が使用できるか。
不快さ	h. 不快感を与える発言・態度(注2)：相手を不愉快にさせるような表現・態度が見られるか。（記述式回答による）

評価は上記の３つのタスクからなるテストの録画ビデオを視聴し、評価項目a～gについては７つの観点別項目それぞれに５段階尺度で、評価項目hについては記述式回答で評価を行った。評価値の尺度は、数値のみではなく、記述文によりその値の意味を示した。評価値３を中程度の値とし、数値が増すほど誤用・不自然な点のない、望ましい値となるように設定した。

３．３．テストの信頼性

テストを行った後、テストの信頼性を検討するため、庄司（1996）に従い、評価者をテスト項目と見なし、クロンバックα係数を求めた（表２・３）。

文法・発音に関しては、やや信頼性係数が低いものの、全体的には一般に必要とされる0.8の値を満たす結果となった。

表２　テスト１における評価者間の信頼性係数（クロンバックα係数）

	語彙	文法	発音	沈黙	反復	発話量
値	0.894	0.760	0.861	0.942	0.882	0.933

表３　テスト２における評価者間の信頼性係数（クロンバックα係数）

	語彙	文法	発音	沈黙	反復	発話量	文の型
値	0.954	0.878	0.821	0.977	0.851	0.965	0.923

４．発話データと評価との関係

上記８つの評価項目のうち、記述式回答による「不快さ」を除いた７つの項目a～gについて、以下に示す客観的な指標に基づいて数値化して発話データを求め、受験者の発話と評価との相関を分析・検討した。

4．1．発話データ

ビデオ録画した発話データを書き起こし、文字化資料を作成した。次に、量的に扱うことができるように、文字化資料を形態素を単位とするデータに処理した。ただし、その目的は、発話量や反復数、発音の不適当な箇所等を数量的にとらえるためであるので、学習者の認識状況から（1）助動詞類は「し・なければ・なり・ません」「ません・でした」のように、（2）あいさつことばは、そのままひとかたまりで「すみません」、「ありがとうございました」のように扱った。

4．2．発話データと指標

評価の観点について、それぞれの指標となる数値を表4のように求めた。発音については、純粋な客観的指標を求めることが難しいため、石崎（1999）を参考に評価者とは別の日本語母語話者教師3名（うち1名は筆者）のうち、2名が不適切とした箇所を数えた。

表4　観点別の評価内容と指標

	項目	指　　標
正確性	語彙	語の選択の誤用箇所数及び既習日本語語彙部分の英語化の箇所数／発話時間（分）
	文法	形態的・統語的誤用箇所数／発話時間（分）
	発音	不適切箇所数／発話時間（分）
流暢性	沈黙	沈黙箇所数／受験者がターンを握っていた時間（注3）（分）
	反復	反復箇所数／発話時間（分）
	発話量	総語数（総形態素数）
	文の型	発話文ごとに文の構造を数値化（1. 単語レベル…3.短文レベル…5. 複文レベル）し、文構造の平均値を求めた。

4．3．指標と評価との相関

発話データと評価との関係について、石崎（1999）や小河原（1996）では相関を見る際に、評価者の評定値を平均化し、その値を使って相関を求めているが、その方法だと、実際の評価者のばらつきの様子は図表からは読み取ることができない。また、標準偏差が提示されていても、それについてはあまり検討されていないことが多い。そこで、本稿では、それぞれの項目について受験者8名の指標値と12名の教師の評価値との相関、すなわち、各項目について8名の受験者×12名の教師評価からなる96ケースの相関の分析を行った。

指標と評価との相関を表５に示す。「沈黙」、「反復」については、テスト１・２ともに１％水準で有意であり、「文の型」についても高い相関係数が得られた。「沈黙」については発話がされているか否かということから、その判断は「有」「無」の二者択一であり、感覚的にもとらえやすく、実態と一致しやすいものと思われる。「反復」は、「沈黙」ほど実態と一致していないが、ある程度的確に反応している。「沈黙」ほど値が高くないのは、教師という性格上、ある程度の反復には常日頃から慣れており、寛容であるという姿勢が反映しているのかもしれない。「文の型」は、教師であれば、文法項目あるいは文型が把握されており、単文か複文かの判断は容易に可能であると思われ、その結果が表れたものと考えられる。

「語彙」・「文法」・「発音」「発話量」についてはどちらか片方のテストでは有意であったが、２回のテストとも同じ傾向とはならず、相関関係は安定しない。これらの原因については、偶然性、テストの内容、評価者、テストの時期等が考えられるが、信頼性係数もある程度あることから、ほかの事柄を見ている可能性も考えられる。そこで、これら４項目について、異なる視点から分析を加えた。

表５ 観点別評価と指標との相関

	テスト1	テスト2
語彙	延べ 0.101 異なり 0.082	延べ 0.514**異なり -0.536**
文法	延べ 0.056 異なり 0.030	延べ 0.370**異なり -0.390**
発音.	延べ 0.200 異なり -0.121	延べ 0.113　異なり 0.293*
沈黙	- 0.581**	- 0.790**
反復	0.377**	0.395**
発話量	0.248*	0.113
文の型	— (注4)	0.640**

(* : p<0.05, ** : p<0.01, N=96)

５．相関が弱かった項目についての再分析

５．１．語彙

図１・２は表５に示した相関について、バブル・チャートを用い示したものである。○の大きさは人数の多さを表す。相関係数はテスト１で r=-0.1 程度、テスト２で r=-0.5 程度であるが、評価値のほとんどが評価値２、３の間に集中してお

り、評価に高低の差があまり見られないことがわかる。このことは、語彙に関して評価を下すほどの明確な特徴が得られなかったことを示していると思われる。

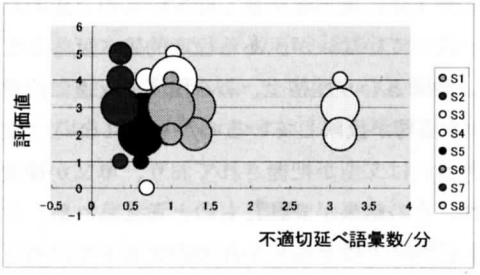

図１　テスト１　１分間あたりの不適切延べ語彙数と評価

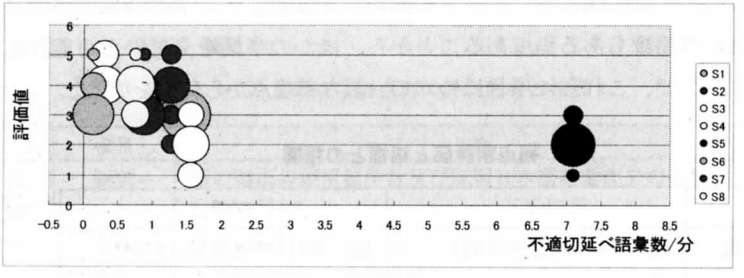

図２　テスト２　１分間あたりの不適切延べ語彙数と評価

語彙に関する評価は「適切に語が使われているか」ということが主眼であったが、「適切な使用」といった場合、「適切なところで適切な語彙を」というように語彙の豊富さを視野に入れていることも考えられる。この点について、発話時間に関係なく、正用の異なり語彙数を指標とする値と評価との相関を求めると、テスト１では r=0.341、テスト２では r=0.306 となった。このことから、評価者は語彙の豊富さにも反応していることがわかるが、基準尺度である「不適切な語彙の頻度」について無視しているとは考えにくく、それらを統合的に見ている可能性もあるため、(「異なり語彙数」－「不適切な異なり語彙数」) という指標についても分析した。その結果、テスト１・２とも同程度の有意な相関 r=0.383, r=0.462 (図３・４) が得られ、評価者は語彙の豊富さに注目しつつ、不適切な語彙も評価の要素に加味し、２つの要素を統合的に評価していることが推察される結果となっ

た。これらのことから「語彙」の項目に対し、教師は「正確さ」だけでなく「流暢さ」にも関心を向けていることがわかった。

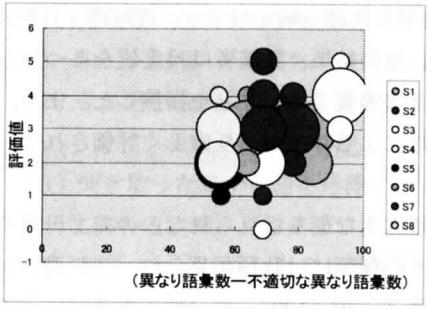

図3 テスト1
(異なり語彙数—不適切異なり語彙数) と評価

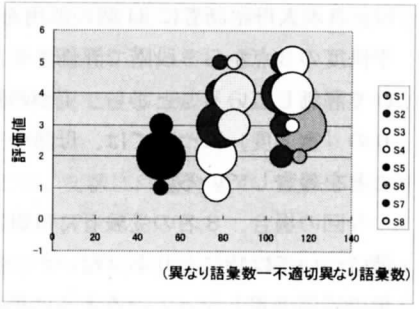

図4 テスト2
(異なり語彙数—不適切異なり語彙数) と評価

5．2．文法

文法については、形態的誤用・統語的誤用数と評価との相関を求めた。テスト2では、r=-0.38 程度と1％水準で有意となったものの、テスト1では、r=0.05 前後と相関は見られなかった（表5）。図で見ると、テスト1では、誤用回数には差があるものの、評価の値は「3」に集中していることがわかる（図5）。テスト2も一見するとテスト1と同様に見えるが、S3（図中右端）の評価を順当に低く答えた評価者がいたため、その分相関係数は高くなったと考えられる（図6）。事実、S3の値を抜いて相関を求めると、r=0.202 と値は低くなり、有意ではなくなる。

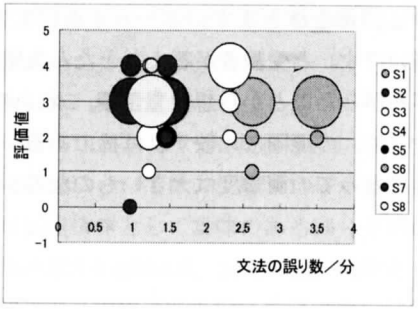

図5 テスト1 1分間あたりの文法の誤りと評価

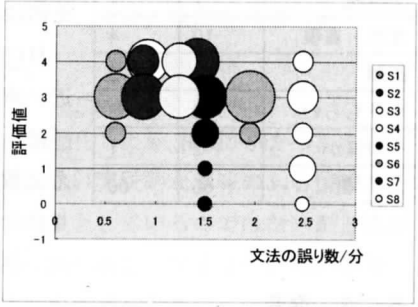

図6 テスト2 1分間あたりの文法の誤りと評価

この結果から、誤用の回数と評価には関係がない、つまり、１回の誤用が質的に同等ではなく何らかの重み付けがあることがわかる。この点については、発話を対象にしたものではないが、趙（1997）の研究でも同様の結果が出ている。趙は、日本人母語話者に84個の誤用を提示した読み物を読ませ、理解度・自然度・不快度の３点から５段階で評価させた。その結果、評価者は尺度値を４つ以上用いて評価していることから、１つの誤りは等質でないことを指摘した。また、誤りの「重要度」については、母語の干渉による誤りがより厳しく評価されていることを報告している。

今回の場合、８名の受験者には母語に大きな偏りは見られないので、母語の干渉については特にとりあげないが、誤りの種類には数種類見られ、この違いが誤用の評価の差となっているように思われる。趙（1997）の分類を参考に２回のテストをまとめたのが表６である。ここでは11に分類したが、この分類から、どの程度の範囲で文に影響を与えるかが誤用の重み付けの一因となっていることが推測される。助詞のような間違いであれば、「サッカーを上手でした」の「を」を「が」に代えるだけで済むが、「町はシーフードです（「町にはシーフードがたくさんあります」という意味で使用）」といった構文に関わる誤りの場合には、瞬時に話者の意図を理解するのは難しいと思われる。また、もう一つ考えられる誤用の重みの要因は、それがどの程度教師にとって聞きなれた誤用かということである。これはビデオ視聴評価時、「そこで行く」「研究室でいます」は聞き流していたのに対し、「海を泳ぎます」「ボールを遊びます」「目に痛いです」と受験者が答えたとたん失望に近い声があがったことから想像できる。これらは、「に」や「で」の混同に比較すれば稀であるがゆえに教師にとっての衝撃度は大きいものなのかもしれない。

表6　文法における誤用

	テスト1	テスト2
助詞	33	31
品詞取り違え	4	8
活用	6	5
テンス	9	2
アスペクト	1	2
指示詞	0	1
文接続	1	1
文型・表現	10	4
語順	0	2
やりもらい	0	3
そのほか	2	1
計	66	60

５．３．発音

発音の不適切箇所数と評価については、テスト１・２とも相関は非常に弱かった。このことから、評価者は発音においても不適切な箇所の数をすべて等質に判

断し、評価していないことがわかる。

ここで、判定者が不適切箇所を指摘する際に述べたコメントに注目してみると、以下のような４つのレベルの理解度があることが判明した。

（1）受験者の意図する語を完全に理解でき、どう不適切なのか明確に指摘できる。：　例）「アクセントが違う」「巻き舌になっている」

（2）音が聞き取れ、文脈にも適合する語が推測できる：例）「５月？９月？（正）５月」「失敗？（正）心配しています」

（3）聞き取った音から語を推測するものの、文脈には適合せず、意味不明のまま終わる：例）「ガイセツ？（正）ガイダンス」：

（4）聞きとれず、まったく推測不能：例）　「わからない」「聞き取れない」

表７　誤用のレベルと数

レベル	テスト１		テスト２	
	個数	%	個数	%
1	61	81.3	64	91.4
2	2	2.7	4	5.7
3	3	4.0	0	0.0
4	9	12.0	2	2.9
計	75	100.0	70	100.0

この４つのレベルは、聞き手の推測不可能度という点から（1）から（4）の順に不適切の重度が大きくなっている。それぞれの尺度の間にどの程度の距離があるのか、尺度の幅は４つでよいのか等まだ明確ではない点もあるが、試験的に上から順に１、２、３、４倍の重みをつけ、不適切箇所の値を計算しなおし、評価との相関をみた。その結果、テスト１で r=-0.330、テスト２で r=-0.309 と相関は有意に強くなった。今回の場合、聞き手の推測可能度から分類したため、イントネーションや単音レベルの誤りもすべて「１」と計算したことに問題があると思われる（表７）。この点については、佐藤（1995）が示すような単音と韻律等との不自然さの評価に関する分析等ともあわせ、今後、検討が必要である。

５．４．発話量

表５では、フィラーや反復等もすべて含めて数値化したが、相関が弱かったため[注5]、発話量については、受験者自らが発話した新しい実質的な意味のある情報あるいは伝達情報として意味のある部分を評価の対象としているのではないかと推測し、総語数から感動詞、反復部分、試験官発話の繰り返し、フィラー、英語部分、聞き取り困難部分を除いた語の数と評価値との相関をみた。

その結果、相関係数はテスト１・２ でそれぞれ r=0.442、r=0.449 となった（図７・図８）。表５の数値より相関がかなり強くなったが、「発話量」が「沈黙」と

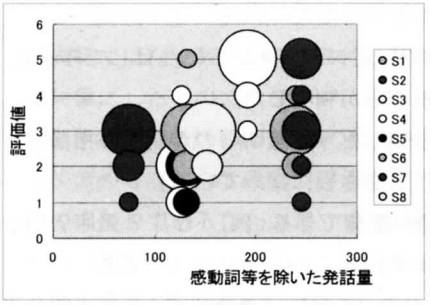

図7 テスト1 感動詞等を除いた発話量と評価

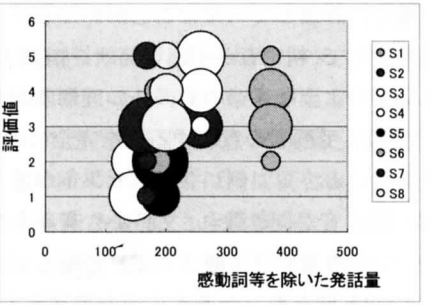

図8 テスト2 感動詞等を除いた発話量と評価

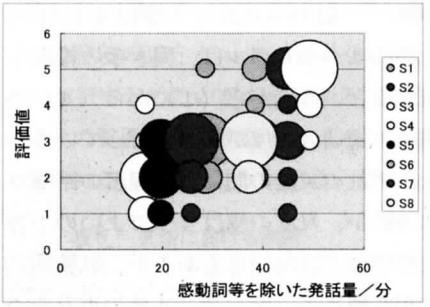

図9 テスト1 1分間あたりの感動詞等を除いた発話量

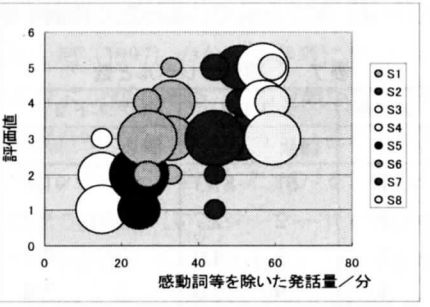

図10 テスト2 1分間あたりの感動詞等を除いた発話量

相対する性質を持つことを考えると、「沈黙」同程度の強い相関を示すことが期待される。また、発話量が少ないにも関わらず高い評価を得ている受験者の存在もあることから「発話量」には決まった時間の中での量が関与することが推察された。そこで、1分間あたりの数値を求め、スピードが加味されているかどうかを検証したところ、相関は、テスト1・2それぞれで r=0.648、r=0.674 となった（図9・10）。この結果より、日本語母語話者教師は、「発話量」について、一定の時間において自発的に使用された、情報として意味のある日本語の量を評価しているということが判明した。つまり、「発話量」にはスピードと量の2つの要因が含まれていることになる。

6．まとめ

研究留学生を対象にした口頭テストにおける観点別の評価項目について日本語

教師の評価と実際の発話との関係を相関分析により調べた。その結果、「沈黙」及び「反復」については、2回のテストとも1%水準で有意であり、これらについては発話に忠実に反応していることがわかった。また、単文か複文かの「文の型」についても、教師は日頃から文型・構文については注意を払っているものと思われ、相関係数の値は高かった。「発話量」については、最初の分析では相関は弱かったが、幾度かの見直しにより、一定の時間の中で発話者自らが語った情報として意味のある部分のみを評価の対象としていることがわかった。「沈黙」や「発話量」に関しては、量的性質を持つため、現象が「有」か「無」の区別をするといったとらえやすさがその要因と考えられる。

一方、「文法」「発音」については、別の視点から分析を加え、これらには1つの誤用を等質に見ることができない重み付けという質的問題が関与していることが明らかになった。「文法」については、文に影響を与える範囲、教師の誤用に対する慣れといったことが評価に大きく関与していることが示唆された。「発音」については発話データを作る際の判定者のコメントから聞き手の推測可能度が影響していることが確認された。ただし、その度合いや詳細な分類についてはさらなる分析が必要である。また、「語彙」については、「不適切な語彙がどの程度現れるか」という基準のほかに、「どの程度語彙が豊富に使えるか」といった要素も含まれていることがわかった。

今回の分析結果より、量的性質が強いものについては、項目が複数であっても発話をそのまま投影した評価が可能であることがわかった。この点については、かなりの信頼性・確実性をもって学習者に評価を提示できるだろう。しかし、質的性質が大きな影響を与えている「発音」「文法」といった項目については、頻度を基準にした尺度では評価とデータとの間にズレが生じることが判明し、評価の方法に改善の余地があることが示された。重み付けの問題も含め、質的な項目を評価としてどのように扱っていくべきなのかについては今後の課題としたい。

付記

本稿は、平成17年度東北大学大学院文学研究科修士論文（仁科 2006）の一部、及び2006年第26回日本語教育方法研究会での発表原稿に、その後の研究成果を加え、加筆修正したものです。ご教示・ご協力くださった方々に感謝致します。

注

（1）テストは日本語コースの中で定期的に行われるテストを想定している。

（2）「不快感を与える発言・態度」とは、状況や場面にふさわしくない言語行動・非言語行動から相手が不快に感じるものを指す。先生に対する「ごめん」、舌打ち等。

（3）ここでは発話が始まってから次の話者の発話が始まるまでを1ターンとした。

（4）テスト1では、初級前半を終えた段階であり、構造面では尺度を設けるほどの幅は見られないと判断した。

（5）阿野(2002)は英語による口頭テスト実験を行い、発語をすべてカウントしているが、発語数と、言い直し・繰り返しの数には正の相関があることを述べている。

参考文献

阿野幸一 (2002)「高校生英語学習者の発話における流暢さと正確さの関係」STEP Bulletin, 14, 39-47.

石崎晶子（1999）「学習者の言語行動に対する母語話者の評価―主観的評価と客観的評価の関係」『第二言語としての日本語の習得研究』3, pp.19-35.

小河原義朗（1996）「韓国人日本語学習者の日本語破擦音の発音と聴き取りの関係について」『東北大学文学部日本語学科論集』6, pp.13-22.

加納千恵子・長谷川守寿・酒井たか子・小林典子 (2004)「日本語Ⅰ」『筑波大学留学生センター日本語教育論集』19, pp.89-95.

佐藤友則（1995）「単音と韻律が日本語音声の評価に与える影響力の比較」『世界の日本語教育』5, pp.139-154.

庄司惠雄 (1996)「日本語研修コースのための口頭能力修了試験」『日本語教育』91, pp.108-119.

鈴木秀明（2004）「教師の教育経験は学習者の発話評価にどのような影響を及ぼすか」『言語科学研究 神田大学大学院紀要』10, 87-106

因京子・栗山昌子・高橋友子 (1991)「研修生コースにおける口頭試験」『九州大学留学生教育センター紀要』3, pp.123-142.

趙南星（1997）『韓国人日本語学習者の誤りについての評価の研究』博士学位論文（東北大学）

Douglas, D.(1994) “Quantity and Quality in Speaking Test Performance.” *Language testing*. 11, pp. 125-144.

―元東北大学大学院生、山形大学―