# 短時間の評価トレーニングが教師の発話評価に及ぼす効果

鈴木　秀明

# The Effects of Short-Term Assessment Training on the Evaluation of Japanese Learners' Speech

Hideaki Suzuki

教育現場において教師が学習者の発話を適切に評価することは極めて重要であるという観点から、本研究では教師の評価技能を高めることを目的として、現職の日本語教師40名に30分の評価トレーニングを行った。トレーニングでは、ACTFLスピーキングガイドラインをもとにした評価基準を応用した講義と、録音された学習者発話を聞いて実際に評価する実技練習を行った。トレーニング前後に、事前テスト・事後テスト方式で、学習者発話評価タスクと評価者の内省コメントにより収集したデータを比較分析した。その結果、教育経験の多少に関わらず、いずれのグループにおいても標準偏差は大きくなり、トレーニング後の内省コメントでは、トレーニングを受けた後に混乱が生じたという報告が多数見られた。これらの結果から今回のトレーニングで使用した教材、指導内容、トレーニングの実施者の指導技術などが評価技能を向上させるには適切ではなかったと考えられる。以上の反省点をもとに、今後はトレーニングの際に、トレーニング内容を実施回数や時間を増やすといった量的なものと、指導方法や教材といった質的なものをともに改善し、さらなる研究の実施が望まれる。

***教師教育　*発話評価　*技能の向上　*教育経験　*トレーニングの質と量**

## はじめに

言語教育に携わる教師には、カリキュラム作成や教材開発をはじめ様々な技能が求められるが、学習者の能力を適切に評価することも教師が習得しなければならない技能の一つである。教師の評価技能を向上させるには、教師の自己研鑽に加え教育機関で評価に関する研修会やワークショップを実施することが望ましいが、物理的および経済的な状況を考えるとこれらの実現は難しく、教

師自身に委ねられているのが日本語教育界の現状である。しかし、教師の評価技能は自己研鑽だけで果たして向上するものなのであろうか。評価研究では、学習者の能力を評価する際の教師の評価基準や、一般人の評価基準を調査したものが数多く見られ（Hadden 1991, Okamura 1995, 中川・石島 1998, 渡部 2002）、教師と一般人の評価に差があると言うことは分かっている。ただし、これらの結果からは教育経験が豊富になるにつれて評価技能が自然に向上するということは報告されていないため、教育経験を積むだけで評価技能が促進されるとは言い切れない。そこで、本研究では教師に専門的知識の習得を目的として、調査過程で短時間かつ簡便な評価トレーニングを実施し、トレーニング前後で教師が学習者の発話を評価する際に、評価の仕方や教師の意識がどのように変化するのかということを明らかにしていく。

## 1. 背景

教師が学習者の発話を安定的にかつ正確に評価するには、言語知識はもとより教授法やテスティングに関する知識も必要となってくる。そこで、以下では本研究に関係のあるスピーキングテストと評価トレーニングに関する先行研究を見ていく。

ブラウン(1999)では、テストを実施する際に、環境、実施方法、受験者自身、採点方法、テスト本体など多くの過程で測定誤差が生じる可能性があると述べ、テストの実施者がどの段階でどの程度の測定誤差が生じるかを十分に認識し、測定誤差を最小限に留められるよう努力することが非常に重要であると示唆している。また、McNamara(1995) および Skehan(1998) では、スピーキングテストに関わる要素として、評価者(rater)、評価基準(scale criteria)、発話(performance)、タスク(task)、参加者(candidate)、潜在的能力(underlying competence)、対話者(interactants)、タスクの質(task qualities)、タスクの条件(task conditions)、第 1 言語と第 2 言語を使い分ける能力(ability for use dual-coding) 等を挙げている。

スピーキングテストの信頼性に関する先行研究では、評価者間での信頼性に関して述べたものが複数見られる。庄司 (1995) では、ACTFL-OPI を参照し

て留学生の口頭能力試験を改訂したものを日本語コースの修了試験として実施し、結果を報告している。それによると、改訂前の口頭能力試験に比べ改訂後の試験では全ての項目において評価者間の信頼性が高くなったと述べている。評価者間の信頼性が高まった要因としては、評価で用いた測定表を具体的にして評価の際の着目点が明らかになったことや、測定表の項目内基準を３段階から５段階にしたことで評価の客観性が増したこと等を挙げている。また、安井 (1994) でも、発話を評価する際に評価方法を検討し、評価表や評価項目内の基準を具体的なものにしたことや、測定の際にビデオテープとオーディオテープを併用した結果、評価の信頼性と客観性が高まったとしている。

学習者の発話を安定的に評価するには、教師に評価技術の向上が求められるが、そのためには評価トレーニングが効果的であると考えられる。そこで、評価トレーニングに関する先行研究についても見て行く。評価トレーニングに関する研究としては、トレーニング前後での学習者評価の変化、教師自身への評価の変化、評価者間の信頼性に関するものが挙げられる。Govoni(1999) では、評価トレーニングが教師の授業内容や教材選択においてどのような効果があるのかについて調査している。実験では、OPI を使用した評価トレーニングを受けた教師と受けていない教師間で教授法や教材選択にどのような違いが見られるかを比較分析した。その結果、評価トレーニングを受けた教師は授業内での一方的な講義や説明が減り、学習者とのコミュニケーションの量が増えた点で有意差が見られたとしている。また、調査後のインタビュー結果から、評価トレーニングを受けた教師はトレーニングを受けていない教師に比べて、学習者に実生活のコミュニケーションタスクに必要な能力を習得させることを熟慮してカリキュラムを組んだり、授業を行ったりしていることも明らかになった。

OPI を使用した評価トレーニングと教師の発話評価の関連性に関するものとしては、Glisan(1998) が実験を行っている。OPI テスターになるためのワークショップを受講していない教師でも学習者の発話の習熟度を正確に予測できるかどうかについて調査したところ、ワークショップの受講経験がない教師は、受講経験のある教師に比べ、学習者の習熟度の過剰判定や過少判定の程度がかなり大きかったと述べている。この結果から学習者の発話を正確に予測するためには OPI を使用した評価トレーニングが必要であると示唆している。評価

者間の信頼性に着目した Halleck(1996) の調査では、中級から超級までの各レベルの OPI の発話テープを OPI のテスター資格保持者とテスターの資格を取得するために現在 OPI のワークショップを受講している大学院生に評価させて、習熟度ごとに評価者間の信頼性を測定した。その結果、超級 (Superior) と中級の中 (Intermediate-Mid) では、グループ間の一致率は高く、信頼性の高い結果が得られたが、上級の上 (Advanced-High)、上級 (Advanced)、および中級の上 (Intermediate-High) では、グループ間の一致率が低くなった。グループ間の信頼性が低くなった要因として、トレーニング受講生は習熟度を適切に判定するだけの明確な知識や判断技術を習得していなかったために、レベルの境界にあたる発話を上のレベルにするのか、下のレベルにするのかで判定の際に困惑していたと分析している。この他にも発話の評価の仕方に関する研究では、Halleck(1992) が OPI の中級、上級、超級の各レベルのインタビューテープを、OPI のテスター資格保持者にどの習熟度に該当するか評価させ、その習熟度に判定を下した根拠をコメントさせたものがある。実験の結果、テスター資格保持者は習熟度を判定する際にはどの習熟度においても、文法等の言語的正確さよりも、機能、内容、コミュニケーションストラテジーの使用等を中心に評価していたと報告している。そして、OPI で習熟度を判定する際には、文法的正確さよりもコミュニケーション能力を重視しているとしている。

日本語教育における評価トレーニングに関する研究としては牧野 (1991) と米田 (2000) が報告されている。牧野 (1991) では、OPI のワークショップを受講することは、テスターの資格の取得に関わらず日本語教師にとっても、カリキュラムの目標設定、教授法、教科書や教材の選択、テスト法等様々な面で思考転換を促してくれる点で効果があるとしている。また、米田 (2000) でも、OPI のワークショップが日本語教育を進める上でどのようなメリットがあるのかを調査するために、ワークショップ受講後に受講生にアンケート調査を実施した。その結果、「会話力のテストとして使える」「会話・口頭能力の捉え方、評価法がわかった」「授業目標の設定にも使える」「教室活動が広がった」など幅広い点で OPI のワークショップが効果的であるというコメントが得られたとしている。

また、教育現場において教師全体が共通認識を持つことは授業内容やカリ

キュラムの作成にも反映されるとした西川他 (2003) は、教師間の評価に関する共通の解釈を涵養することを目的として発話評価のワークショップを実施した。調査者の所属する日本語学校で会話テストを担当している 20 数名の教師を対象に、会話テストを実施する前に 7 回ワークショップを開催した。会話テストの担当者は自身の都合の良い時に任意でワークショップに参加し、評価基準に関するトレーニングを受講した。ワークショップでは、会話テストで録音された中級及び上級学習者の発話テープを材料として使用した。これらの発話テープは事前に OPI テスター資格を所持している 4 人の調査者によって予め採点され、評価基準の確認と評価基準点が決定されていた。参加教師は発話テープを聞きながら、学校独自の採点用評価メモシートに採点した。その後、調査者からワークショップの正答と評価基準に基づいた解説が行われ、参加教師からの質疑応答にも応じる時間を設けた。なお、1 回あたりのワークショップに要した時間は 1 時間程度で、各回で聞いたテープの本数は 1 本ないし 2 本だった。ワークショップを実施する際には、評価基準において意見が分かれる項目に関しては、評価しやすい典型的な会話例を取り出し解説するなど、より具体的にするなどの工夫をしたり、参加教師が学生の会話レベルの全体像を把握するために担当レベル以外のテープも聞かせるような工夫をしたともしている。結果として、ワークショップを三回程度受け、発話テープを 3 本以上評価する経験を持つと、教師間での評価基準も揃い始め、会話テストを採点する際にも教師の発話評価が安定してくると報告されている。

以上の先行研究から、教師は評価トレーニングの受講によって、程度の違いはあるものの、専門的知識を徐々に習得していくことが分かった。しかしながら、評価トレーニングの効果とトレーニングを受講する教師の教育経験との関係については、明らかにされていない。そこで、本研究では、短時間の評価トレーニングの実施は、教師が非母語話者の発話を評価する際にどのような効果をもたらすのかということについて、現職の日本語教師を対象に調査する。

## 2. 調査

### 2.1 調査目的

ここでは、スピーキングテスト、評価トレーニングに関する先行研究で明らかにされた事実に基づいて、本研究の質問を提示する。本研究は、以下の質問に答えるべく計画され実施された。

質問：短時間の評価トレーニングは、教師が非母語話者の発話評価を行う際にどのような効果を及ぼすか。また、トレーニング効果は教育経験によって異なるか。

なお、本研究では、教育経験 5 年以上の教師を教授年数の多いグループ、教育経験 5 年未満の教師を教授年数の少ないグループと定義する。

### 2.2 参加協力者

参加協力者は、日本語教師 40 人（女性 38 人・男性 2 人）で、教師の教授年数は 10 ヶ月から 13 年まで、平均教授年数は 5.6 年であった。全員が現在日本国内で教授活動に携わっている。以下では、この参加協力者のことを参加教師と呼ぶ。

### 2.3 調査で使用したインタビューテープ

調査では、牧野他 (2001) の ACTFL-OPI の中級と上級のインタビューテープから 5 つのサンプルインタビューを評価対象のテープとして使用した。インタビューテープを選択する際に、本調査では発話相手から受ける影響を少なくするために、ロールプレイを除外した。また、このインタビューテープを選択した理由の一つとして、ACTFL-OPI のインタビューテープは、既に OPI テスター資格保持者によって発話者の言語習熟度が分析されていて、どのレベルに該当するのかが明らかにされていることを付け加えておく。

### 2.4 本研究における評価トレーニング

本研究では、調査過程で独自の評価トレーニングを実施した。この評価トレー

ニングは講義と練習の二部で構成されていて、トレーニングの前半では評価基準に関する講義を行い、ACTFL Provisional Revised Speaking Guideline(1998)に基いて作成した評価基準の資料を参加教師に配布し、調査者自身が講師役となり評価項目および評価基準について説明した。ACTFL-OPIの評価基準を参考に評価項目を決定し、テキストタイプ、文法、音声、語彙、流暢さ、機能、内容の七項目を設定した。各項目ごとにそれぞれの概念及び具体的な内容を説明した後に、評価項目と習熟度の関係について説明を行った。また、評価基準を説明する際には具体例を挙げ、できるだけ参加教師にわかりやすく伝わるように努めた。続いてトレーニングの後半では、実際に非母語話者の発話テープを参加教師に聞かせ、発話テープが初級から超級までのどの習熟度に当てはまるのかを項目ごとに評価させる実践的な練習を実施した。なお、講義の間、参加教師には資料を見ながらメモをとることは許可したが、資料内容や評価基準に関する質問は受けつけなかった。また、練習の際に参加教師には講義の際の資料を見ながら習熟度を判定してもらったが、評価後に調査者からのフィードバックは時間の制約のために特に行わなかった。これらの点で参加教師に制限を加えたのは、トレーニングで参加教師が得る知識量を全員揃えるためである。以上、評価トレーニング全体に要した時間は講義と練習を合わせて約30分であった。

### 2.5　評価タスク

評価タスクは、評価トレーニング前後の分析的評価、教師の自己基準の記述、評価トレーニングに対する内省コメントの3つを実施した。トレーニング前後の分析的評価では、参加教師に非母語話者のインタビューテープをテキストタイプ、文法、音声、語彙、流暢さ、機能、内容の7項目について項目ごとの評価を課し、参加教師にテープに録音された発話を聞かせ、評価シートに項目ごとに5段階で評価させた。この5段階評価の基準は、ACTFLのガイドラインの習熟度分類に対応するよう設定し、1と2が初級、3が中級、4が上級、5が超級とした。なお、今回の調査ではトピックの影響を避けるためにカウンターバランスをとり、トピックと評価の順番の組み合わせを考えて実施した。40人のグループを10人ずつ4グループに分け、それぞれ評価するトピック

の順番を入れ替えた。また、4 グループを分ける際には、グループ間の教授年数が均等になるように配慮した。

教師の自己基準では、教師が何に着目して発話能力を評価しているのかを調べるために、参加教師にインタビューテープの発話を評価する際に 5 段階評価でどれにしたかという理由を評価シートに記述させた。また、この質問とは別に日頃の教授活動で発話を評価する際の基準についても合わせて質問し、記述形式で回答させた。

評価トレーニングに対する内省コメントでは、トレーニングの前後で教師が非母語話者の発話を評価する際にその評価の仕方や意識に変化が生じたかどうかを調べるために、調査の最後に内省コメントを記述させ、参加教師にトレーニング前後で評価の着目点が変わったかどうかを質問し、「変化した」、「少し変化した」、「あまり変化しなかった」、「よくわからない」の選択肢から選ばせた。さらに、「変化した」、「少し変化した」を選択した参加教師には、評価の仕方や評価の際の意識がトレーニング後にどのように変化したのかをできるだけ詳しく記述させた。

### 2.6　手順

調査は 2002 年の 7 月から 8 月にかけて神田外語大学大学院研究室及び都内にある語学教育機関の会議室において、調査者自身が立会いの元で行った。調査では、まず目的と手順について口頭説明を行った。その後、教師の自己基準、分析的評価（中級テープ、上級テープの順）、評価トレーニングに対する内省コメントの順に評価タスクを行った。なお、調査の過程で参加教師に関する背景アンケートも合わせて実施した。調査にかかった時間は全体で 1 時間 30 分程度であった。

### 2.7　分析方法

調査で得られた分析的評価の測定結果は、教育経験の多少のグループごとに集計して、平均点と標準偏差を出した。また、測定結果について一般化が図れるかどうかを調べるために統計処理を行った。調査で得られたデータを統計処理する際には、統計処理ソフトの StatView を使用した。教師の自己基準に関

する記述データは、ACTFL-OPIの測定基準をもとに調査者が10項目に分類した。評価タスクに対する内省コメントデータは、調査者がコメントの種類ごとに7種類に分類し、教授年数の多少のグループごとにどのような傾向があるのかをまとめた。自己基準に関するデータおよび内省コメントのデータは、グループ全体および教授年数の多少のグループごとに集計した。

## 3. 結果と考察

### 3.1　評価トレーニング前後の分析的評価

評価トレーニング前後の分析的評価においては、教授年数の少ないグループのみに特定のテープの種類と評価項目で評価平均点に変化が見られた。この点を具体的に見ると、教授年数の少ないグループが上級テープを評価した際に、7項目の中で文法において、トレーニング後に評価平均点が低くなることが明らかになった(表1)。一方、教授年数の多いグループでは、中級テープ、上級テープのいずれにおいても、7項目全てでトレーニング前後での評価平均点に違いは見られなかった。教授年数の少ないグループでどのような変化が起きたかという点については、後の内省コメントの分析の際に詳しく述べることにする。

表1．教授年数の少ないグループのトレーニング前後における上級テープの評価平均点

| 評価項目 | 測定時期 | | t 値 | P 値 |
|---|---|---|---|---|
| | トレーニング前 | トレーニング後 | | |
| テキストタイプ | 4.71 | 4.43 | 1.55 | n.s |
| 文　　法 | 4.33 | 3.86 | 2.50 | P<0.02 |
| 音　　声 | 4.10 | 3.71 | 1.71 | n.s |
| 語　　彙 | 4.05 | 3.81 | 0.87 | n.s |
| 流　暢　さ | 4.05 | 4.05 | 0 | n.s |
| 機　　能 | 4.19 | 4.00 | 0.70 | n.s |
| 内　　容 | 4.05 | 3.81 | 0.96 | n.s |

### 3.2　評価トレーニング前後の評価のバラつき

評価トレーニング前後における標準偏差の変化について調べるために、教育経験（年数多い、年数少ない）、テープの種類（中級、上級）、測定時期（トレーニング前、トレーニング後）を独立要因として、繰り返しのある三要因の分散分析を行った。その結果、テープの種類 ($p<0.01$) と測定時期 ($p<0.01$) で有意差が見られ、テープの種類と測定時期が標準偏差の変化に関係していることが明らかになった（表 2）。また、教育経験・テープの種類・測定時期の 3 要因間の交互作用では、わずかに有意に至らなかったものの、やや関係が見られた ($0.05<p<0.06$)。

表 2. 教育経験とテープの種類と測定時期によるバラつきの変化

| 変動要因 | 自由度 | 偏差平方和 | 平均平方 | F 値 | P 値 |
|---|---|---|---|---|---|
| 教育経験（A） | 1 | 0.03 | 0.03 | 1.37 | n.s |
| テープの種類（B） | 1 | 0.29 | 0.29 | 14.76 | 0.008 |
| A×B | 1 | 0 | 0 | 0.16 | n.s |
| グループ内 | 24 | 0.47 | 0.02 | | |
| 測定時期（C） | 1 | 0.11 | 0.11 | 11.30 | 0.003 |
| A×C | 1 | 0.02 | 0.02 | 2.23 | n.s |
| B×C | 1 | 0.02 | 0.02 | 2.38 | n.s |
| A×B×C | 1 | 0.04 | 0.04 | 4.16 | 0.053 |
| C×グループ内 | 24 | 0.22 | 0.01 | | |

この点について、教育経験、テープの種類、測定時期がどのように影響しているのかを調べるためにさらに分析したところ、表 3(a)(b) に示す結果が得られた。教育経験と測定時期との関係について見ると、トレーニング前後におけるグループごとの標準偏差の変化では、教授年数の少ないグループの変化 (+0.13) は、年数の多いグループの変化 (+0.05) に比べて 2.6 倍大きかった。テープの種類と測定時期との関係について見たところ、トレーニング前後におけるテープの種類による標準偏差の変化は、上級テープの変化 (+0.13) のほうが、中級テープの変化 (+0.05) に比べて 2.6 倍大きかった。

表 3(a). 教育経験によるトレーニング前における標準偏差

| 測定時期 | | トレーニング前 | トレーニング後 |
|---|---|---|---|
| 教授年数 | 多い | 0.70 | 0.75 |
| | 少ない | 0.70 | 0.83 |

表 3(b). テープの種類によるトレーニング前後における標準偏差

| 測定時期 | | トレーニング前 | トレーニング後 |
|---|---|---|---|
| テープの種類 | 中級 | 0.79 | 0.84 |
| | 上級 | 0.61 | 0.74 |

教育経験とテープの種類と測定時期の関係について分析した結果、トレーニング前後における教育経験とテープの種類による標準偏差の変化を示した表 4 によると、教授年数の多いグループはトレーニング前後での変化はテープの種類に関係なく小さかった ( 中級テープ +0.06、上級テープ +0.04) が、教授年数の少ないグループではテープの種類によって標準偏差の変化が異なり、特に上級テープを評価した際に標準偏差がかなり増加したことが明らかになった。(中級テープ +0.03、上級テープ +0.22)。

表 4. グループごとのトレーニング前後における標準偏差の増加幅

| テープの種類 | | 中級テープ | 上級テープ |
|---|---|---|---|
| 教授年数 | 多い | +0.06 | +0.04 |
| | 少ない | +0.03 | +0.22 |

### 3.3　評価トレーニング後の内省コメント結果

今回の調査の結果、教授年数の少ないグループでトレーニング前後の分析的評価で評価平均点に変化が生じた。この事実から、教授年数の少ないグループではトレーニング後に文法項目内の評価基準の量や質に変化があったと推測される。また、教育経験とテープの種類と測定時期の関係について分析した結果、

トレーニング後はいずれのグループも標準偏差値が増加した。この点は、トレーニング後に評価者間の評価に差が激しくなったことを意味している。以上の点を詳しく分析するためには評価トレーニング後の内省コメントと照らし合わせてみることが必要である。評価トレーニングに対する意識変化（表 5）では、教授年数の多いグループでは 19 人中 14 人(約 74%)がトレーニング後に何らかの変化を感じていた。これに対して、教授年数の少ないグループでは 21 人中 19 人（約 90%）がトレーニング後に変化を感じていたことが判明した。

表 5.　評価トレーニング後の意識変化

| 項目 | 教授年数 | |
|---|---|---|
| | 多いグループ（n＝19） | 少ないグループ（n＝21） |
| 変化した | 3 | 7 |
| 少し変化した | 11 | 12 |
| あまり変わらない | 5 | 2 |

そこで、この意識変化を詳細に観察するために評価タスクに対する内省コメントデータを内容別に分類し、7 種類に分けたものが表 6 である。

表 6.　評価トレーニング後の内省コメントの項目と内容

| | |
|---|---|
| 1．評価項目数の増加 | 評価基準の対象となる項目が増えた |
| 2．項目内基準の詳細化 | 項目内の評価基準量が増え詳細になった |
| 3．評価項目内の基準と習熟度の意識化 | 項内の基準と発話の習熟度の対応を認識 |
| 4．評価基準の意識化 | 主観的評価から評価基準に基づいた評価へ |
| 5．全体的評価から分析的評価へ | 全体的評価から項目ごとの分析的評価へ |
| 6．評価の際の混乱（1） | 講義内容は理解できたが、評価の際に混乱 |
| 7．評価の際の混乱（2） | もとの自己基準が影響し、評価の際に混乱 |

このコメント結果をグループ別に見ていくと、教授年数の少ないグループでは評価トレーニング前は評価基準が明確になっていない状態で漠然と評価していたが、トレーニング後は評価の際の着目点が明確になり、評価項目内の基準も詳細になったというコメントが見られた (表 7)。すなわち、トレーニング前は発話を主観的に評価していたが、トレーニング後は評価基準に照らし合わせる形式に変化したため、評価平均点が低くなったものと考えられる。

表 7. 年数の多少によるトレーニング後の内省コメント

| コメントの種類 | 教授年数 | |
|---|---|---|
| | 多いグループ | 少ないグループ |
| 評価項目数の増加 | 4 | 7 |
| 評価項目内基準の詳細化 | 4 | 6 |
| 評価項目内の基準と習熟度 | 4 | 4 |
| 評価基準の意識化 | 8 | 11 |
| 全体的評価から分析的評価へ | 1 | 3 |
| 評価の際の混乱（1） | 1 | 10 |
| 評価の際の混乱（2） | 5 | 0 |

ここまでの結果をまとめると、教育年数の少ないグループではトレーニング前後で中級テープの文法において評価平均点に変化が生じ、また評価の際の意識変化も見られた。この事実から、今回の評価トレーニングを受けた結果、以前の主観的で曖昧だった評価基準が整理され、意識化された可能性がある。ただし、これは中級レベルだけで、他の習熟度には当てはまらないものである。また、他の評価項目にも言えない。

今回の調査結果から、教育経験の違いに関わらず、どちらのグループでもトレーニング後に標準偏差が大きくなり、評価者間の評価のずれが広がったことが明らかになった。この点についてグループごとに比較したところ、経験年数の多いグループに比べて、経験年数の少ないグループの方がかなりバラつきの増加幅が大きいことが分かった。また、グループ内でテープの種類とバラつきの変化の関係を見たところ、中級テープよりも上級テープでバラつきが非常に大きくなっていた。以上の事実から、グループ間で程度の違いはあるものの、評価トレーニング後にどちらのグループでも評価者間のバラつきが大きくなったことが判明した。

この標準偏差の増加を考察する上で、トレーニング後の内省コメント(表7)から、多くの参加教師が今回の評価トレーニングによって評価の際に混乱を起こしていたことが確認された。以下ではグループごとに内省コメントを分析していく。

このコメント結果によると、「評価の際の混乱」に関するコメントでは、「評価の際の混乱（1）」においては、教授年数の多いグループでは 1 人しかコメントしていなかったのに対して、教授年数の少ないグループでは 10 人がコメントしていた。したがって、教授年数の少ないグループは年数の多いグループに比べて評価トレーニングで評価項目や評価基準をインプットされたものの、実際の場面で使いこなせるようにならなかった割合が高いことが推測される。一方、「評価の際の混乱（2）」では、教授年数の多いグループでは 5 人がコメントしていたのに対して、年数の少ないグループでは誰もコメントしていなかった。すなわち、年数の多いグループでは評価トレーニングを受けた結果、既に評価者自身が持っている評価項目や評価基準を新しい基準に当てはめようと試みたが、適切に調節できなかったと思われる。

この事実をより具体的に確認するために両グループの特徴を示す参加教師のコメントを取り上げて考察する（表 8）。以下は、今回の評価トレーニングの後に評価タスクを行った際に、どのように評価の混乱が起こったのかについてグループごとにコメントされたものの一部である。

表 8. 評価トレーニング後の意識変化の内省コメント例

| 参加教師グループ | 内省コメント |
|---|---|
| 教授年数少ない | 表で評価の基準の説明を受けたが、それだけでは自分が判断する際の適格な判断力まではつかない。具体的にテープと照らし合わせて、「ここではこんな間違いをしているから、この人は○○の項目については×級です。」というトレーニングが必要かと思った。（参加教師 No.4）<br><br>トレーニングによって様々な角度から評価するということや、その分類まで習ったので考え直したい。しかし今日の評価基準がすぐに身につくとは思えない。（参加教師 No.36） |
| 教授年数多い | トレーニング後に新たな観点も取り入れて評価しようと思ったが、全ての項目に適切に答えるのは難しいと感じた。（参加教師 No.11）<br><br>トレーニング前は自分の企業内の評価基準を中心に評価していた。トレーニング後は新たな評価基準に合わせて評価を試みたが、少し混乱した気がする。（参加教師 No.29） |

これらの内省コメントを見ると、教授年数の少ないグループの参加教師No.4およびNo.36は、今回の評価トレーニングによって項目内の基準がより詳細になり、新たに評価基準に照らし合わせる評価形式に変化していったことがうかがえる。しかし、これら教授年数の少ないグループの参加教師は、評価トレーニングで知識は得たものの、実際の場面でこの基準をすぐに運用できるようにならなかった。この点については、新たな評価基準を身につけるために更なる評価トレーニングの必要性も示唆している。一方、教授年数の多いグループの参加教師No.11およびNo.29は、評価トレーニング前から既に自分自身の中に評価基準があり、それに基づいた評価をしていたことがうかがえる。トレーニング後は新たな評価基準に照らし合わせた評価を試みたが、既存の自己基準に影響され新しい評価基準に適切な調節ができなかったことが読み取れる。

以上のことから、両グループとも評価の際に混乱を起こしたことでは同じだったが、その混乱の原因には違いがあることと考えられる。そして、この混乱の性質の差が標準偏差の増加幅の差に反映したものと思われる。すなわち、教授年数の多いグループは、既にあった自己基準を調整しようとした際に生じた混乱だったために標準偏差の増加幅が小さかったのに対して、教授年数の少ないグループは、トレーニング後に評価基準の量が増え、さらにその基準に当てはめた評価をしようとしたため、標準偏差の増加幅が大きくなったのではないだろうか。

### 3.4　評価トレーニングが参加教師に及ぼした影響

今回の評価トレーニングは、教師が専門的知識を習得することを促進することを目的としてデザインされたが、評価タスクを実施した結果、項目別分析的評価、標準偏差、およびトレーニング後の内省コメントのいずれをとっても、結果的には参加教師にとってマイナスに作用したと言わざるを得ない。この事実から今回のトレーニングは様々な点において問題点があったと思われる。そこで、以下では今回のトレーニングの問題点を量的な側面と質的な側面に分けて取り上げていく。なお、西川他 (2003) で報告されたワークショップの内容と本調査を比較対照しながら述べていく。

まず、トレーニングの量的な側面について見ると、本研究では 30 分の評価トレーニングを 1 回実施し、トレーニング前後で評価の仕方の違いを見るために評価タスクを実施し、参加教師の教育経験の違いを比較した、しかし、参加教師に評価基準を確実に理解して内容を定着させ、さらに評価基準を運用可能にするためには、時間も回数も不十分であったと思われる。この点については、新しい基準を認識したが完全には理解できなかった、時間が足りなかった、再度トレーニングを受けたいなどのトレーニング後に記述されていた内省コメントとも一致する。西川 (2003) では、ワークショップは 7 回実施され、1 回あたり 1 時間の練習を行っていた。そのため、参加回数の少ない教師は評価が不安定だったが、3 回程度受講すると評価が安定するという結果を報告している。この事実を見ても、やはり複数回にわたりワークショップを受講し、継続した評価トレーニングを受けることが効果的であると思われる。

次に、トレーニングの質的な側面について見ると、本研究では、トレーニングの際に講義と練習の二つの形式を取ったが、講義では単に評価基準の説明を一方的に行うだけで参加者と評価基準についての定義や認識をすり合わせる作業は行わなかった。また、練習の際にも発話テープを採点させた後に、正答を知らせて評価基準についての解説等のフィードバックも実施しなかった。そのために、参加教師と調査者の間に評価基準において共通認識が持てなかった可能性がある。さらに、トレーニングの際に聞いた発話テープの本数は 1 本だけであったのも練習としては不適切であったと考えられる。これに対して西川 (2003) では、ワークショップでテープを採点させた後に OPI テスター資格保持者が採点結果と採点基準の解説、参加教師からの質疑応答に答えるなどのフィードバックを毎回行っていた。このフィードバックにより、参加教師の評価基準が徐々に理解され、教師間での共通認識が涵養されていったのではないだろうか。

以上の点から見ると、評価トレーニングを実施する際には使用教材、トレーニングにおける指導方法、トレーニング実施者の熟練度等が効果をもたらすかどうかという点においても非常に重要であると思われる。

## 4. 結論および今後の課題

本研究では、教師が学習者の発話能力を評価する際に必要な専門的知識を促進することを目的として、評価トレーニングを実施した。現職の日本語教師を対象に、ACTFL スピーキングガイドラインを参考にデザインされた短時間でかつ簡便な評価トレーニングを行った。参加教師のトレーニング効果を見るために事前事後テスト方式および評価者の内省コメントを収集した。その結果、教育経験の多少に関わらず、トレーニング後にグループごとの標準偏差の値は大きくなり、特に教育経験の少ないグループが上級テープを評価した際に、標準偏差の値が一番増加したことが明らかになった。また、トレーニング後の内省コメント結果から、参加教師の多くが評価の際に混乱を招き、評価基準が不安定になってしまったことも分かった。

今回の評価トレーニングが結果的に参加教師にとって適切なものにならなかったのは、トレーニングの内容が質的にも量的にも不充分であったためである。調査者は、教育現場で実用可能なトレーニングという視点に基づいて本研究のトレーニングをデザインしたために、トレーニングに関わる多くの要素について十分に吟味しないまま調査を実施してしまった。しかし、実際に今回のトレーニングを実施して、評価トレーニングに関連する要素はトレーニング期間、トレーニング回数、1 回あたりの時間数、使用教材、トレーニングにおける指導方法、トレーニング実施者の熟練度、トレーニングを受ける側の評価技術など多岐に渡ることを学んだ。したがって、トレーニング効果を測定するためには、これら関連する条件を十分に認識した上で、何を変数にするのかということを明確にした上で研究をデザインすることが不可欠である。今後は、この点に留意して更なる研究を実施したい。

[謝辞]
* 本稿は 2004 年 8 月に日本語教育国際研究大会で発表したものを加筆修正したものである。堀場裕紀江先生には、論文執筆及び修正の折に何度も重要なご指摘とご助言を頂いた。また、小林美代子先生には、査読の際に示唆に富んだコメントや温かい励ましを頂いた。ここに両名に心から感謝を表する。

[参考文献]

庄司恵雄（1995）「日本語研修コースのための口頭能力試験」，『日本語教育』91 号，pp108 － 119.
中川道子・石島満沙子（1998）「会話の上達度を計る評価基準」，北海道大学留学生センター紀要第 2 号，pp169 － 184.
西川寛之・西部由佳・山中都・山辺真理子（2003）「パフォーマンス・アセスメント中心の口頭表現能力テスト　―比較可能性の保証を高める評価者の涵養―」日本語教育学会春季大会予稿集，pp183 － 188.
バックマン・F・R（1997）『言語テスト法の基礎』，みくに出版
ブラウン・J・D（1999）『言語テスト法の基礎』，大修館書店
牧野成一（1991）「ACTFL の外国語能力基準およびそれに基づく会話能力テストの理念と問題」，『世界の日本語教育』第 1 号，pp15 － 32.
牧野成一・鎌田修・山内博之・斎藤眞理子・荻原稚佳子・伊藤とく美・池崎美代子・中島和子（2001）『ACTFL - OPI 入門―日本語学習者の「話す」能力を客観的に測る』，アルク
安井澄江（1994）「会話試験の評価に関する試み」名古屋大学日本語・日本文化論集，pp111 － 139.
米田由貴代（2000）「OPI を授業に生かす―受講者から見た OPI ワークショップのメリット／デメリット」『月刊日本語』4 月号，アルク
渡部倫子（2002）「日本語口頭運用能力をどのように測定するか」，日本語教育学会春季大会予稿集，pp179 － 184.
American Council on the Teaching of Foreign Languages(1986): *ACTFL proficiency guidelines*. Hastings-on-Hudson,NY:ACTFL.
American Council on the Teaching of Foreign Languages(1999): *ACTFL Provisional Revised Speaking Guidelines.* Hastings-on-Hudson, NY:ACTFL.
Glisan,E,W.(1998).Assesing Students' oral proficiency in an outcome-based curriculum: Student performance and teacher institution,*The Modern Language Journal.* 82,1-18.
Govoni,J.M.(1999).Effects of the ACTFL-OPI-Type-Training on student performance, instructional methods, and classroom materials, in the secondary foreign language classroom, *Foreign Language Annals.* 32,No2,189-204.
Hadden,L.H.(1991):"Teacher and Nonteacher Perceptions of Second-Language Communication" *Language Learning*,41-1,1-24.
Halleck,G.B.(1992).The oral proficiency interview: Discrete point test or a measure of communicative language ability?, *Foreign Language Annals.*25,No3,227-231.
Halleck,G.B.(1996).Interrrater reliability of the OPI:Using academic trainee raters. *Foreign Language Annals.*29,No2,223-238.
McNamara,T.(1995):Modelling Performance:opening Pandora's box'. Applied *Linguistics.*16,159-179.
Okamura, A(1995):Teachers' and Nonteachers' Perception of Elementary Leaners' Spoken Japanese,*The Modern Language Journal.*79,29-40.
Skehan,P(1998):*A Cognitive Approach to Language Learning,*Oxford:Oxford University Press.