スピーチにおける生徒相互評価の妥当性
―項目応答理論を用いて―

茨城県立竹園高等学校　教諭　深澤 真

概要　本研究の目的は，スピーチにおける日本人高校生の相互評価にはどの程度の妥当性があるかを検証することである。妥当性は，1）構造的要素，2）一般可能性的要素，3）外的要素，4）内容的要素，5）実質的要素，6）影響的要素の6つの側面より検証された。主な分析方法として項目応答理論のラッシュ・モデル分析を用いた。その結果，生徒相互評価の構造的要素については部分的に妥当性が認められ，その他の5つの妥当性の側面についてはすべてにおいて十分な妥当性が示された。6つの妥当性の要素の検証結果をまとめると，スピーチにおける日本人高校生による相互評価には一定の妥当性があると考えられる。また，生徒相互評価には，教員評価に比べ平均値が高く，標準偏差が低い傾向も見られた。これらの研究結果に基づき，生徒相互評価の活用について教育的示唆を行う。

1 はじめに

現在，コミュニケーション能力の育成は高等学校での英語教育における焦点と言っても過言ではない。「英語が使える日本人」育成のための行動計画やスーパー・イングリッシュ・ランゲージ・ハイスクール（SELHi）の指定も文部科学省によって行われ，コミュニケーション能力の育成はますます強調されている。さらに，新学習指導要領では，「授業を実際のコミュニケーションの場」（文部科学省，2009, p.92）と位置づけており，スピーキングを含めた4技能を有機的に結びつけ，バランスよく学ぶことが求められる。

しかし，コミュニケーション能力の1つとしてスピーキングの能力をどう伸ばすかに重点が置かれている一方で，その評価方法についてはあまり強調されていない（秋山, 2000）のが現状である。スピーキングの能力の評価に消極的になる理由としてまず挙げられるのは，スピーキングの評価は主観的になりやすく（McNamara, 2000），言語テストの分野において評価が最も難しい分野の1つであるということである（馬場他, 1997）。また2つ目の理由として日本の教育現場では多くの場合1クラスに40人程度の生徒がおり，一人一人スピーキングのテストをしていくことが難しい（Heaton, 1975）ことが考えられる。

これらの問題点を少しでも軽減し，スピーキングの評価をより実用的なものにするよう，生徒相互評価の活用を考えた。生徒相互評価を教員評価と併用することにより，多面的かつより客観的に評価が行えることと，評価のための時間と労力の軽減にもつながると考えられる。

本研究では高校生による相互評価を「生徒相互評価」とし，主に項目応答理論を使って生徒相互評価の結果を分析し，その妥当性の検証を行う。

2 先行研究

2.1 妥当性

従来，妥当性とはテストが測ろうとしているものを正確に測っているかどうか（Hughes, 1989）と考えられており，内容妥当性，構成概念妥当性，併存的妥当性，予測妥当性など，それぞれ独立したものとして分類されていた。しかし，今日では，妥当性

は構成概念妥当性を中心としたテストの解釈と使用に関する1つのまとまった概念としてとらえられている（Bachman, 1990）。この妥当性の概念は信頼性も含む多面的なものであり（Chapelle, 1999），さまざまな観点から妥当性を見ていくことが不可欠である。Messick（1996）は，この妥当性の概念を1）内容的要素，2）実質的要素，3）構造的要素，4）一般可能性的要素，5）外的要素，6）影響的要素の6つの側面に分類した。1つ目の内容的要素とは，知識や技能などの測定される内容が評価タスクと関連しているかという妥当性の側面である。2つ目の実質的要素とは，理論的プロセスが実際の評価タスクにおいて行われているかについての妥当性の側面である。3つ目の構造的要素とは，評価の内的構造が構成概念の内的構造と一致しているかの側面を表している。4つ目の一般可能性的要素とは，テストの母集団，設定やタスクによりどの程度テストで得られた得点や解釈が一般化できるかという側面のことである。5つ目の外的要素とは，テスト得点と他の測定結果との関係が予測される関係をどの程度反映しているかという妥当性の側面を示している。6つ目の影響的要素とは，長・短期的に見てテスト得点の解釈と使用における意図した影響と意図しなかった影響がどのようにあるかという妥当性の側面である。本研究でもこのMessickの妥当性の解釈に基づき，研究を行う。

2.2 相互評価とその妥当性

2.2.1 相互評価

相互評価は，一般に継続的評価（定期テストなどのように成績の決定などに用いる一括的評価に対して，家庭学習や課題など継続的に行う評価）の1つとして考えられている（松沢, 2002）。Brown（1998）は，相互評価を「1人かそれ以上の生徒の言語または言語運用力の判断を生徒に求める評価」（p.54）と定義している。

Brown（1998）は，相互評価の長所として評価に余計な時間や手段を必要としないという実用的な面を挙げている。また，相互評価は，生徒の動機づけを高めたり（Orsmond, Merry, & Reiling, 1996），学習目標をはっきり認識させたりする（Luoma, 2004）効果も持っている。

一方，短所として，相互評価が比較的主観的な評価方法であることと，入試のような重要度の高い（high-stakes）状況においては信頼性に問題があることを挙げている（Brown, 1998）。

2.2.2 相互評価の妥当性

総括的評価の一部，つまり成績の一部として相互評価を使おうとする場合，比較的主観的評価とされる相互評価の妥当性を検証する必要がある。Miller & Ng（1996）は，教員評価と相互評価の相関を分析し妥当性の検証を行った。41人の大学生を対象に英語スピーキングの授業2クラスで実験を行い，教員評価と相互評価の間に比較的高い相関（r = .68-.80）を認め，相互評価は教員評価と比較しうると結論づけた。また，相互評価の信頼性を高める条件の1つとして高い英語の能力を挙げている。Hughes & Large（1993）も44人の大学生を対象に，オーラル・コミュニケーションにおける相互評価の妥当性を検証する実験を行った。その結果，教員評価と相互評価の相関は高く（r = .83），相互評価の妥当性が認められた。Stefani（1994）では教員評価，相互評価，自己評価の相対的な信頼性の研究を行い，教員評価と相互評価には非常に高い一致が見られ（r = .89），生徒の評価が教員と同程度の信頼性を持ちうると結論づけた。Stefaniはまた，相互評価の平均値は教員評価の平均値より高く，標準偏差は教員評価と比べて相互評価の方が小さい傾向にあることも指摘した。

さらに，Nakamura（2002）とMatsuno（2009）では，項目応答理論に基づいてラッシュ・モデル分析を行い妥当性の検証を行った。Nakamuraは，オーラル・プレゼンテーションのクラスの大学生12人を対象とした相互評価の妥当性を検証した。無作為抽出された5人の相互評価者のうち1名だけ項目応答理論のモデルにミスフィットした評価者（misfit rater）と判断されたが，他の4人の評価者は評価の厳しさに差はあるものの一貫した評価をしていたことがわかった。その結果をもとに，学生による相互評価は一定の信頼性を備えていると結論づけた。Matsunoも同様に項目応答理論を用いて，大学生91人と教員4人を対象にライティングにおける自己評価，相互評価，教員評価の比較を行った。その中で，多くの相互評価者は評価者内信頼性が一貫しており，自己評価や教員評価と比べてバイアスも少なかった。このことから，相互評価はライティングのクラスである程度役立てることができること

が検証された。

一方で，相互評価の妥当性について否定的な先行研究もある。Orsmond et al.（1996）は，教員評価と相互評価の結果を基に回帰分析を行った。その結果，２つの評価間の相関は高かったが，偏差に有意差が認められ，相互評価の信頼性に疑問を投げかけた。Freeman（1995）もまた，210人の大学生を対象に教員評価と相互評価の分析を回帰分析を用いて行い，その結果それらの評価間の相関は中程度であり，よい発表を低く評価し，よくない発表を高くつける傾向が見られた。この結果は，学生が信頼性のある評価者になることが難しいことを示した。

以上のような先行研究からもわかるように，これまでの相互評価についての妥当性の検証結果は分かれており，とらえようとしている妥当性の側面や分析方法によっても結果に違いが見られた。

2.3　項目応答理論とラッシュ・モデル

項目応答理論とは「ある困難度を持ったテストの項目に，ある能力を持った受験者がどのように応答するか，ということに関して確率的なモデル（probabilistic model）を設定し，それに基づいて受験者の応答データを分析したり，テストを開発したりするための理論」（靜・竹内・吉澤, 2002, p.100）である。このモデルの設定のためには，２つの前提が仮定されており（大友, 1996），１つは「局所独立の仮定」，つまり受験者が１つの項目に正答できる確率は，そのテストの他の項目に正答できる確率に影響を受けないという仮定である。もう１つの前提条件は「一次元性」であり，用いられるテスト項目はただ１つの能力分野を測定するものでなければならないという前提である。項目応答理論の特徴として，1）異なるテストを用いても共通の尺度上で能力測定が可能なこと，2）異なる受験者集団に用いても共通の項目特性に関する値を求めることが可能なこと，3）項目ごとの測定の精度を求めることが可能なことなどが挙げられる（大友）。

項目応答理論の１パラメータ・ロジスティック・モデルは開発者である Rasch の名前からラッシュ・モデルと呼ばれている。１パラメータ・ロジスティック・モデルは「項目困難度パラメータ」を基準にしている（大友, 1996）。さらに，項目困難度，受験者の能力の２つの相（facets）に加えて，評定者の厳しさや課題の難しさなどの相を加えたものが「多相ラッシュ測定」（Many-Facets Rasch Measurement）である（靜他, 2002）。この分析方法は，評価者の主観が介入した測定状況の分析に適しており，スピーキングテストの分析を行った秋山（2000）やライティングの評価者を分析した長橋（2008）でも使用されている。この測定の代表的なソフトウェアが Linacre の開発した FACETS であり，本研究でも使用する。

3　研究の目的・仮説・特徴

3.1　研究の目的

本研究の目的は，スピーキング活動の中のスピーチに焦点を当て，日本人高校生による生徒相互評価にはどの程度の妥当性があるか検証することである。

生徒相互評価は Brown（1998）の定義に基づき（2.2.1参照），１人の生徒のスピーチを発表者以外のクラス生徒全員で評価する。また，評価項目は発音，文法，流暢さ，態度，総合評価の５項目であるが，本研究では総合評価に焦点を当てる。

3.2　研究の仮説

妥当性検証にあたっては，Messick（1996）を基に，妥当性の６つの要素，1）構造的要素，2）一般可能性的要素，3）外的要素，4）内容的要素，5）実質的要素，6）影響的要素の側面より検証を行う。なお，妥当性の要素の順序については，本研究の分析手法ごとに分け，並べ替えた。

また，Chapelle（1999）の妥当性検証の方法に従い，①仮説を作り，②仮説を検証するための関連した証拠を提示し，③証拠と理由づけを提示し妥当性に関する主張を行うものとする。研究の仮説は次のとおりである。

1）構造的要素の検討
　妥当性の仮説１：生徒相互評価の結果は一次元である（具体的には，生徒相互評価のデータは項目応答理論のモデルに適合している）
　妥当性の仮説２：生徒相互評価の信頼性は高い
2）一般可能性的要素の検討
　妥当性の仮説３：クラスにより生徒相互評価の結果に大きな差はない

3）外的要素の検討
妥当性の仮説４：生徒相互評価と教員評価との相関は高い
妥当性の仮説５：生徒相互評価と教員評価では平均値において統計的差異はない
4）内容的要素の検討
妥当性の仮説６：スピーチは授業での学習内容と関連している
5）実質的要素の検討
妥当性の仮説７：生徒相互評価はスピーキングの能力を測っている
6）影響的要素の検討
妥当性の仮説８：生徒相互評価はスピーチのテストによい影響を持つ

3.3 研究の特徴

本研究の特徴は，1）高校生を対象として，2）項目応答理論を用い，3）妥当性を多面的に検証する，相互評価の研究となっている点である。第１に，これまでの相互評価の先行研究はほとんどが大学生を対象としており，高校生を対象とした相互評価の研究はほとんどない。第２に，分析手法として項目応答理論を用いる点である。Nakamura（2002）とMatsuno（2009）を除き，相互評価における先行研究ではほとんどの場合，教員評価と相互評価の相関をとる方法により妥当性の外的要素の側面を中心に検証が行われてきた。本研究では項目応答理論を用いることにより構造的要素の側面も検証している。第３に本研究では，構造的要素，一般可能性的要素，外的要素，内容的要素，実質的要素，影響的要素の多様な観点から妥当性の検証を行っていることが特徴である。

4 研究方法

4.1 参加者

本研究には公立高校の生徒157名（１年生２クラス80名，２年生２クラス77名）が参加した。２年生に関しては2006年に，１年生に関しては2008年に実験を実施した。参加者のうち有効データは，発表者145名，評価者117名，アンケート116名であった。外国籍を有する生徒も５名いるが，英語を母語とはしていない。この高校は研究者の勤務校であり，英語と理科の教育に力を入れている進学校である。また，教員評価者として研究者を含む日本人英語教員４名と英語指導助手（ALT）１名の計５名が参加した。日本人英語教員に関しては高校教員としての経験を10年以上有しており，ALT は勤務３年目である。

4.2 手順

今回の実験は授業（１年生は英語Ⅰ，２年生は英語Ⅱ）の一環として行った。教員評価とともに生徒相互評価の結果も成績の一部に含めることとし，生徒にもその旨を伝えた。まず，事前に熟達度テスト（α = .78）を作成した。この問題は英検（２級～３級）の問題などを参考に作成され，リスニング問題（15問），語彙問題（15問），文法・語法問題（20問）の計50問を30分間で行った。

次に，授業１時間をかけて評価者訓練を行った。生徒はまず各レベルのスピーチの例をビデオで見てから，評価練習（４回），模擬評価（３回）を行った。教員の評価者訓練についても，生徒評価者訓練に先立って同様の手順で行った。評価項目は発音，文法，流暢さ，態度，総合評価の５項目である。また，評価基準表の作成にあたっては，Council of Europe（2001）の Common European Framework の６段階評価（A1～C2）を参考にして日本人高校生のスピーキングの熟達度に合わせて作成した（資料１）。

評価者訓練の次の授業から３時間かけてスピーチ発表と生徒相互評価を行った（各クラス40名程度，１時間当たり13～14名）。スピーチの発表は１人２分間程度で，直後の１分間を使って発表者以外のクラス全員で生徒相互評価を行った。教員（３～４名）もこのときに評価を行ったが，教員の一部が授業に参加できなかった場合は，後でビデオを見て評価を行った。

スピーチ発表と生徒相互評価の後１週間以内に，生徒相互評価に関するアンケート調査を実施した（資料２）。

4.3 分析方法

3.2 の妥当性の６つの要素の具体的分析方法については，小泉（2005）を基にしてまとめた（表１）。

妥当性の構造的要素の側面については，生徒相互評価と教員評価の結果を基にラッシュ・モデル分析とクロンバック α の信頼性係数を使った分析を行う。ラッシュ・モデル分析では FACETS（Linacre,

■ 表 1：妥当性の 6 つ要素と分析手法

要素	■妥当性検証で吟味する側面 ・チェック項目	本研究での具体的分析手法
構造的要素	■テストの内的構造（タスク・項目・設問間の関係）の検証 ・データの次元が，仮定された構成概念の次元と一致（適合）するか	・項目応答理論を使ったラッシュ・モデル分析 ・評価者間信頼性
一般可能性的要素	■時間・グループ・受験状況・タスク・評価者の変化などにより，テストのプロセスと構造が変化するのかの検証 ・得られた得点の解釈が，構成概念の領域に一般化できるか	・項目応答理論を使ったラッシュ・モデル分析
外的要素	■テスト得点とテストの外的構造（他の測定結果・背景的変数）の関係の検証 ・外的基準との関係（関係のなさ）が得点の意味と一致するか	・ピアソンの相関係数 ・*t* 検定
内容的要素	■テスト内容と測定領域が一致しているかの検証 ・内容が関連しているか ・内容に代表性があるか	・アンケートの分析
実質的要素	■受験者が項目・タスクにどう反応しているかの検証 ・専門家でない人にとって，テストが意図した構成概念を測っているように思えるか	・アンケートの分析
影響的要素	■テスト得点を解釈・使用する際に社会的な影響があるかを検証 ・テスト得点を解釈・使用する際に，意図した影響が短期的・長期的にあるか	・アンケートの分析

（注）小泉（2005）を基に作成。

2008）を用いて，求められた Infit Mean Square の値を基に項目応答理論にミスフィットした評価者（misfit rater）を見つけ，その割合を基に生徒相互評価が項目応答理論の前提となっている一次元性を持つかどうかの分析を行う。ミスフィットの基準は，McNamara（1996）に基づき「Infit Mean Square の平均＋ 2×Infit Mean Square の標準偏差以上」をミスフィットとした。ミスフィットしている生徒評価者が全体の10％未満で（Stansfield & Kenyon, 1995），ミスフィットと判断された発表者が全発表者の 2 ％未満なら（McNamara），項目応答理論のモデルに適合し一次元性を有すると判断する。

生徒相互評価の信頼性に関しては，SPSS（2008）を使用し，クロンバックの α 係数を用いて評価者間信頼性の分析を行う。信頼性の基準は，一般的には $\alpha \geqq .80$と考えられている（三浦, 2004）が，Lado（1961）は，スピーキングの評価の場合は $\alpha \geqq .70$でも十分であるとしている。本研究では Lado に基づき評価者間信頼性は，α 係数 .70以上を高い信頼性とする。

一般可能性的要素の分析も FACETS を使って行い，ミスフィットの基準については構造的要素の分析同様 McNamara（1996）の基準を使う。

妥当性の外的要素の側面の分析には相関と *t* 検定を使う。生徒相互評価と教員評価との相関についての分析は，ピアソンの積率相関係数を使って分析を行う。相関の基準に関しては，清川（1990）に基づき（ほとんど相関がない＜±.20，±.20 ＜低い相関＜±.40，±.40 ＜中程度の相関＜±.70，±.70 ＜高い相関），高い相関を $r = .70$以上とする。

生徒相互評価と教員評価との平均値の比較は，*t* 検定を行い 5 ％水準で平均値の有意性の検定を行う。

妥当性の内容的要素，実質的要素，影響的要素の分析と判断の基準については「5 結果と考察」の中で説明する。

5 結果と考察

5.1 構造的要素

まず，生徒相互評価の妥当性を構造的要素の側面より検討する。生徒相互評価の構造的側面については，1）項目応答理論のミスフィット項目の観点と2）評価者間信頼性の観点の 2 つの観点から検討する。

5.1.1 FACETS による分析概観

次ページの図1は項目応答理論統計ソフトFACETS で生徒相互評価の結果を分析した図である。左端の列は，ロジット（logit）と呼ばれる項目応答理論で使われる単位であり，この図では0が平均を示している。第2列は項目困難度を示し，ここでは生徒のスピーチの能力を表している。上に行けば行くほど高い能力を示し，下に行くほど能力は低くなる。第3列は，個々の評価者の厳しさを表しており，上部に位置している評価者は評価が厳しく，下部に位置している評価者は甘い評価者と言える。ロジットは理論的にはプラス・マイナス無限大まで表すことができるが（大友・中村, 2002），－3.0～3.0の範囲で表されることが普通である（大友, 1996）。したがって，かなり厳しい評価者が1名（最も上部にいるA）とかなり甘い評価者が3名（最も下部にいるA，C，D）いたことがわかる。A，B，C，Dのアルファベットは生徒評価者の所属するクラスを表しており，T は教員評価者を示している。第4列は，A～D のクラスと教員のグループとしての評価の厳しさを示している。A，B が 2 年生のクラスであり，C，D は 1 年生のクラスである。第5列は，6段階評価の尺度を表している（C2が最高，A1が最低）。この尺度と第2列のスピーチの能力をあわせて見てみると，発表者のほとんどが6段階評価中B1～C1の評価を受けており，評価 A2と評価 C2だった生徒はそれぞれ数名であり，評価 A1の生徒はいなかったことがわかる。

5.1.2 生徒相互評価の一次元性

FACETS による分析に基づき，項目応答理論のミスフィットの観点から以下の仮説を検証する。

妥当性の仮説 1 ：生徒相互評価の結果は一次元である（具体的には，生徒相互評価のデータは項目応答理論のモデルに適合している）

もし，項目応答理論のモデルに生徒相互評価が十分に適合しているのであれば，項目応答理論の前提である生徒相互評価の一次元性（生徒相互評価が同じ 1 つの能力を測っていること）が満たされると考えられる（McNamara, 1996）。セクション4.3の基準に基づき，ミスフィットの基準値，全体の中の割合，信頼性をまとめたものが表 2 である。

■表 2：ミスフィットの基準値・割合と信頼性

	Infit Mean Square の *M*（*SD*）[ミスフィットの基準値]	ミスフィットの割合（基準を超えた数／全体）	信頼性 [Separation]
評価者（*N* = 122）	.98（.38）[1.74]	5.74%（7/122）	.94 [3.90]
発表者（*N* = 145）	1.00（.29）[1.58]	4.83%（7/145）	.99 [8.13]

（注）*M* = 平均値；*SD* = 標準偏差；*N* = 人数

評価者の Infit Mean Square の平均値は.98で，標準偏差は.38なので，ミスフィットの基準は.98＋2×.38，つまり Infit Mean Square の値1.74以上をミスフィットとした。その結果，ミスフィットした評価者は 7 名（クラス A，2 名；クラス B，3 名；クラス C，1 名；クラス D，1 名；T（教員），0 名）であった。全体の評価者は122名であったので，ミスフィットした評価者の割合は全体の5.73%であった（うち生徒評価者は117名でその中のミスフィット評価者の割合は6.00%）。

一方，発表者のミスフィットの基準は1.00＋2×.29 = 1.58となり，Infit Mean Square の値が1.58以上をミスフィットとした。ミスフィットした発表者は 7 名（クラス A，2 名；クラス B，4 名；クラス C，0 名；クラス D，1 名）で発表者全体の4.83%であった。

以上の結果を項目応答理論の前提である一次元性の基準（ミスフィットの評価者は全体の10%未満，ミスフィットの発表者は全体の 2 %未満）にあてはめてみると，ミスフィットした評価者の割合は全体の5.73%（生徒評価者中の割合は6.00%）で基準を満たしたが，ミスフィットした発表者の割合は全体の4.83%で基準を満たさなかった。したがって，妥

▶ 図 1：FACETS 分析結果（Valuable Map）

Logit	+Speaker	-Rater	-Rater (class)	Scale
9				(C2)
	•			
8	•			
	•			
	•			
7	*			
	•			-----
6				
	•			
	*•			
5				C1
	**			
	***•			
4	**•			
	*•			
	***	A		-----
3	**			
	***	D D D		
	**			
2	***	A [A1] C C		
	****•	A A [A2] C C D D		B2
	****	T A A B D D D		
1	*****•	A A A B [B1] C C D		
	***•	T A A B C D D D		
	***	T A [B2] B B B B B C C C C C C D D	D	
0	****	A A A A A [B3] B B C C C C D D D D	A C T	-----
	***	T A B B B C C C C D [D1] D	B	
	**	A A A A B B B C C D D D D		
-1	**	A B B B C [C1] D		
	**	T B C C D D		
	**•	A B B B B C		B1
-2	**	A A A B		
	**•	C		
	*•			
-3				
	*•	D		-----
	•	A C		
-4	•			
-5				A2
	•			
	•			
-6				(A1)
Logit	•＝1 ＊＝2	-Rater	-Rater (class)	Scale

（注）＊は 2 人の発表者を表し，・は 1 人の発表者を表している。□はミスフィットした評価者を表している。

当性の仮説 1 である生徒相互評価の一次元性は部分的に支持された。

5.1.3 生徒相互評価の信頼性

次に，評価者間信頼性の観点から，以下の仮説を検証する。

妥当性の仮説 2 ：生徒相互評価の信頼性は高い

生徒相互評価と教員評価の評価者間信頼性は表 3 のとおりである。Lado（1961）のスピーキングにおける高い信頼性の基準 $\alpha \geqq .70$と比べると，生徒相互評価の評価者間信頼性の数値は $\alpha = .91$〜 .99とどのクラスにおいても例外なく非常に高い値となっている。このことから生徒相互評価は高い信頼性を有すると判断することができ，妥当性検証の仮説 2 は強く支持された。

■ 表 3：評価者間信頼性

評価	クラスA	クラスB	クラスC	クラスD
生徒相互評価	.99 (n = 30)	.91 (n = 28)	.99 (n = 30)	.98 (n = 29)
教員評価	.93 (n = 3)	.89 (n = 3)	.92 (n = 4)	.90 (n = 4)

（注）n = 評価者の人数

5.1.4 構造的要素の検討

項目応答理論のモデルに適合し一次元性を持つかという観点では生徒相互評価はミスフィットした評価者が全体の10％未満という基準は満たしたものの，ミスフィットした発表者が 2 ％未満という基準を満たさなかったので，生徒相互評価の結果が一次元性を持つことは部分的に支持された。

一方，評価者間信頼性については生徒相互評価の結果は高い信頼性を有することが確認された。これら 2 つの結果を総合すると，生徒相互評価は構造的要素の面からは妥当性が部分的に認められたと考えられる。

5.2 一般可能性的要素

生徒相互評価における妥当性の一般可能性的要素の側面についての検討は，次の仮説を検証することにより行う。

妥当性の仮説 3 ：クラスにより生徒相互評価の結果に大きな差はない

生徒相互評価の結果をクラス単位で比較し，スピーチという同じタスクを使いながら，発表する生徒・評価する生徒が異なる各クラスの結果に大きな差がなければ，得られた得点の解釈がより一般化できると考えられる。また，クラス A，B とクラス C，D は学年も異なるため，学年による差がなければさらに一般化できると思われる。

仮説 3 の検証にはラッシュ・モデル分析を用いた。その結果が表 4 である。検証の方法は，各クラスの評価結果がフィットしているかどうかで行う。ミスフィットの基準はセクション 5.1.2で用いた基準とすると1.01＋(2×.10) = 1.21で，Infit Mean Square の値1.21以上をミスフィットとした。

各クラスの Infit Mean Square の範囲は，一番高い値のクラス A で1.06，一番低いクラス C で.87であった。Infit Mean Square の基準となる値は1.00であり，各クラスともそこを中心に大きなぶれはなく，ミスフィットの基準である1.21を超えているクラスはなかった。評価の厳しさについては，最も評価の厳しいクラスが D（.26），評価の甘いクラスが B（-.41）であった。 2 つのクラスの Logit の違いは.67で，図 1 でもわかるように厳しさの面においてもクラス間で大きな違いがなかったことがわかる。さらに，公正化された平均得点でもほとんどばらつきは認められず（SD = .06），得点面からもクラスによる違いはほとんどなかった。

学年の観点（クラス A，Bは 2 年生；クラス C，Dは 1 年生）から分析結果を見てみると，Infit Mean Square において基準となる1.00を中心に 2 年生が1.00以上（クラス A = 1.06；クラス B = 1.04），1 年生が1.00以下（クラス C = .87；クラス D = .94）という傾向は見られるが，いずれも 1 より大きく離れていないので学年の違いと言えるような大きな違いは認められなかった。また，評価の厳しさについても学年によるものと見られる明らかな違いはなかった。

これらの結果により，クラスにより生徒相互評価の結果に大きな差はないという仮説 3 は支持され，一般可能性的要素の側面からも生徒相互評価の妥当性が示された。

■ 表 4：クラスとしての評価結果分析

評価者（クラスとして）	実測			公正化された平均得点	Logit（厳しさ）	Infit MnSq
	総得点	評価数	平均			
A（*n* = 30）	4175	1052	4.00	3.89	.11	1.06
B（*n* = 28）	3727	952	3.90	4.03	-.41	1.04
C（*n* = 30）	3878	1024	3.80	3.95	-.12	.87
D（*n* = 29）	4337	1102	3.90	3.85	.26	.94
T（*n* = 5）	1926	509	3.80	3.88	.16	1.16
M	3608.6	927.8	3.90	3.92	.00	1.01
SD	868.2	214.9	.10	.06	.24	.10

（注）*M* = 平均値；*SD* = 標準偏差；*n* = 人数；T = 教員評価者

5.3　外的要素

外的要素に関しては，生徒相互評価と教員評価との比較により検討を行う。教員評価との比較には相関と平均値を用いる。相関を使用するのは，教員評価と同じような観点で優れたスピーチには高い評価を，あまりよくないスピーチには低い評価をしているかどうかを検証するためである。また，平均値についても検証を行うのは，仮に同じような観点で評価していたとしても評価の厳しさに差がある場合，評価結果が異なることがあるからである。そこで，教員評価と比べて同じ観点で評価し，評価の厳しさにも差がなければ教員評価と近い評価をしていると考えられるので外的要素の検討には相関と平均値の比較を併用する（Fukazawa, 2007）。

5.3.1　教員評価との相関

まず，1つ目の教員評価と生徒相互評価の相関の観点から，次の仮説の検証を行う。

妥当性の仮説 4 ：生徒相互評価と教員評価との相関は高い

表 5 は生徒相互評価の平均値と教員評価の平均値の相関を示している。ピアソンの積率相関係数は .79 から .93（$p < .01$）であった。クラスD の相関が他のクラスに比べやや低いが，清川（1990）の基準に基づくと相関係数はいずれも .70 を上回っており，生徒相互評価と教員評価の間には高い相関があると考えられる。したがって，仮説 4 も支持された。

■ 表 5：生徒相互評価と教員評価の相関

	クラスA（*n* = 36）	クラスB（*n* = 35）	クラスC（*n* = 35）	クラスD（*n* = 39）
相関	.92**	.93**	.90**	.79**

（注）*n* = 人数　**$p < .01$

5.3.2　教員評価との平均点の比較

次に，生徒相互評価と教員評価の平均値の観点から，以下の仮説を検証する。

妥当性の仮説 5 ：生徒相互評価と教員評価では平均値において統計的差異はない

t 検定を使って，生徒相互評価と教員評価間の平均値を比較した。表 5 で示されているように，クラス A〜D のいずれのクラスにおいても 5 %水準で平均値に有意差は見られなかった。したがって，平均値において生徒相互評価は教員評価と統計的な差異はないと結論づけられ，仮説 5 も支持された。

■ 表 6：*t* 検定の結果

クラス	*t* 値	*p* 値
A（*n* = 36）	-9.38	.35*
B（*n* = 35）	-1.90	.06*
C（*n* = 35）	-1.03	.31*
D（*n* = 39）	-.74	.94*

（注）*n* = 人数　*$p < .05$，両側検定

5.3.3　外的要素の検討

このセクションでは相関と平均値の観点より，生徒相互評価と教員評価の比較を行った。相関ではクラス D において他クラスよりやや低かったものの，

どのクラスにおいても生徒相互評価と教員評価の高い相関が認められた。クラスDについて記述統計（資料3）では特に他のクラスと大きく異なる点は見られなかった。また，セクション5.1.2でも，クラスDでミスフィットしていた生徒評価者は1名，ミスフィットしていた発表者が1名と相関に大きな影響を与える要素とは考えづらいと思われる。

また，平均点において，統計的有意差は見られなかったものの，クラスBの生徒相互評価の結果は，他クラスと比べて教員評価の結果との平均値の違いが.40と最も大きかった（教員平均＝3.51；クラスB平均＝3.91）。ただし，差が0.5を下回っているため評価のカテゴリーが変わる可能性は大きくないと予想される（例えば，教員評価とクラスBの平均値を実際のカテゴリーに置き換えるとどちらもB2と判断される）。

以上の結果より，相関と平均点において生徒相互評価は教員評価と非常に似た評価パターンをとっていると考えられる。このことから，妥当性の外的要素の面からは生徒相互評価は十分な妥当性が認められる。

5.4 内容的要素と実質的要素

生徒相互評価の妥当性における内容的要素と実質的要素の側面に関しては，アンケートの分析を基に次の2つの仮説を検証する。

妥当性の仮説6：スピーチは授業での学習内容と関連している

妥当性の仮説7：生徒相互評価はスピーキングの能力を測っている

表7は生徒相互評価についてのアンケート結果をまとめたものである。質問12と13についてアンケートの参加者が少ないのは，2008年に実験の行われたクラスCとクラスDのみに実施したためである（質問13については欠損値があったため人数が1名減っている）。質問は主に1～5までの選択肢から選ぶ形式のアンケートで，記述回答も含まれる。1は「とてもそう思う」で賛成を表し，4が「全くそう思わない」で反対を表している（5は「わからない」）。5を除外した1～4の中央値2.5以下を肯定ととらえることとする。内容的要素の側面についての質問12と13では，いずれの回答の平均値も2.5以下となり生徒はスピーチの発表と評価項目のどちらも授業におおむね関連があったと考えていると思われる。また，研究者以外の教員評価者にも同じアンケートを実施したところ，同じ質問に対する回答の平均値が生徒よりさらに下回り，授業との関連性をより強く感じていたことがわかった。これらの結果から仮説6は支持されていると認められ，生徒相互評価は内容的要素の面からも妥当性を有していると考えられる。

■表7：生徒相互評価に対する生徒の反応

質問事項	*M*	*SD*
Q.12　内容と授業の関連性（*N* = 59）	1.87（1.67）	.62
Q.13　評価項目と授業の関連性（*N* = 58）	1.78（1.50）	.66
Q.18　相互評価の適切性（*N* = 116）	1.85	.53

（注）*M* = 平均値；*SD* = 標準偏差；*N* = 人数；（　）教員評価者による結果

次に，実質的要素について検討する。生徒が相互評価は発表者のスピーキングの能力を測っていると感じているかどうかは質問18で尋ねた。この質問に対して生徒は肯定的にとらえており（*M* = 1.85），仮説7も支持された。

アンケートの結果では，各質問に対する回答の平均値は1.78～1.87で2に近い回答となっている。つまり，内容的要素・実質的要素いずれもどちらかというと肯定的にとらえていることがわかった。これらのことから，内容的要素，実質的要素のどちらの側面についても生徒相互評価の妥当性がおおむね認められたと考えられる。

5.5 影響的要素

生徒評価妥当性の影響的要素についてもアンケートの分析を行い，次の仮説を通して検討する。

妥当性の仮説8：生徒相互評価はスピーチのテストによい影響を持つ

実験後行ったアンケート（資料2）の質問30で生徒が生徒相互評価から学んだ点について自由記述回答を求めた。その結果を肯定的，中立，否定的の3つのカテゴリーへ分類したものが表8である。この

分類は研究者が行った。

■ 表 8：相互評価の効果

肯定的	中立	否定的
62	3	2

（注）回答人数 = 67

生徒相互評価に関するアンケートに答えた生徒116名中，回答した生徒は67名であった。そのうち相互評価について肯定的回答は62名であった。肯定的にとらえていた生徒のうち，特に話すことと聞くことの領域について生徒相互評価は参考になったと答えた生徒が多かった。その他の肯定的な意見としては，目標となるスピーチを見つけられた，評価することが自分のスピーチの参考になった，他の人の意見が聞けてよかった，自分に足りないものを見つけたなどの意見であった。中立的意見では「人前で立ってスピーチすることの大変さ」という意見があった。このコメントについては，大変さを学んだと肯定的にとれる一方，大変だったと否定的にもとれるため中立とした。さらに，否定的なコメントとして，自分の行った相互評価の客観性に疑問を持つ意見などがあった（資料 4 ）。

相互評価を行った116名の生徒のうち，自由記述回答において62名（53.4%）がなんらかの形で生徒相互評価について肯定的効果を認めており，仮説 8 は支持されたと考える。したがって，影響的要素の面からも相互評価の十分な妥当性が認められた。

5.6　全体的考察

全体考察では，1）ミスフィットした評価者と発表者，2）生徒相互評価の評価者間信頼性，3）教員評価と比べた生徒相互評価の平均点と標準偏差の 3 点について述べる。

第 1 に，ミスフィットした評価者についてなんらかの傾向が見られるか，ラッシュ・モデル分析の予期せぬ反応（unexpected response）と生徒の熟達度の 2 点について検討した。表 9 がミスフィットした評価者の予期せぬ反応数をまとめたものである。予期せぬ回答数は多い人で 5，少ない人で 1 であり，ミスフィットしていなかった他の評価者と比べても大きな違いは見られなかった。また，評価の厳しさの点でも厳しすぎる予期せぬ応答数（11）と甘すぎる予期せぬ応答数（10）はほぼ同数で，ミスフィットした評価者に共通する傾向は見られなかった。

■ 表 9：予期せぬ反応数

ミスフィットした評価者	予期せぬ応答数	厳しすぎ	甘すぎ
A1	3	1	2
A2	4	2	2
B1	4	2	2
B2	1	0	1
B3	2	2	0
C1	2	2	0
D1	5	2	3

また，ミスフィット評価者と英語熟達度の関係をまとめたのが表10である。英語熟達度による違いを見るため，熟達度テストの結果を基に 1 ～ 4 の 4 グループ（ 1 が最上位で 4 が最下位）に分けた。 4 グループ間に違いがあるかを見る際に，一元配置分散分析を用いた。等分散性が満たされなかったため Brown-Forsythe の修正を行い，その結果どのグループ間にも有意差が確認された（p = .00*）。Miller & Ng（1996）は，英語力の高さを相互評価の信頼性を上げる 1 つの条件としていたが，本研究ではどの熟達度グループにもほぼまんべんなくミスフィットした評価者がおり，熟達度との関係は特に見られなかった。

■ 表10：ミスフィット評価者と英語熟達度の関係

英語熟達度グループ	ミスフィット評価者数
1（n = 30）	2
2（n = 32）	2
3（n = 29）	2
4（n = 26）	1

（注）n = 人数；英語熟達度グループ 1 は最上位，4 が最下位

一方，ミスフィットした発表者にはいくつかの傾向が見られた。 1 つは 7 名のミスフィットした発表者のうち 6 名は 2 年生（クラス A，B）から出ており，1 年生（クラス C，D）は 1 名であった（5.1.2 参照）。また，表11のように発表者を英語熟達度別に 4 グループに分けたとき［グループ間の違いについては一元配置分散分析を上述と同じ手続きで行い，有意差が確認された（p = .00*）］，ミスフィットした発表者はどちらかというと下位に位置していることがわかった。この結果をただちに一般化する

のは難しいが，少なくとも本研究では１年生より２年生の方が，また成績上位者より下位者の方がミスフィットする発表者が多いことがわかった。

■表11：ミスフィット発表者と英語熟達度の関係

英語熟達度グループ	ミスフィット発表者数
1（*n* = 35）	0
2（*n* = 32）	2
3（*n* = 40）	3
4（*n* = 37）	2

（注）*n* = 人数；英語熟達度グループ 1 は最上位，4 が最下位

第２に，表３で示されているように生徒相互評価の評価者間信頼性は，どのクラスにおいても教員評価の評価者間信頼性よりも高くなる傾向が見られた。Brown（1996）は，信頼性係数は評価者の人数による影響を受けやすいとしている。評価生徒数は各クラス40名程度なのに対し，教員の人数は３～４名であり，本研究でも評価者の人数の影響を受けたものと考えられる。

第３に，資料３の記述統計を見ると，生徒相互評価と教員評価の結果に２つの傾向が見られた。１つ目は，生徒相互評価の平均点は教員評価の平均点よりも一貫して高い傾向である。つまり，生徒は教員よりもやや甘めに評価をしている傾向がうかがえる。２つ目の傾向は，生徒相互評価の標準偏差は教員評価のそれより小さいことである。この傾向はStefani（1994）の結果とも一致しており，教員がA1～C2の幅をより広く活用している一方で，生徒はA1やC2といった極端に悪かったりよかったりする評価を避けている様子がうかがえる。ラッシュ・モデル分析のカテゴリー統計（表12）でも主にB1～C1の３段階を中心に評価が行われており，A1の評価が特に少ないことでも裏づけられている。

■表12：カテゴリー統計

カテゴリー	使用回数	割合（%）
A1	28	0.6%
A2	350	7.5%
B1	1174	25.3%
B2	1892	40.8%
C1	945	20.4%
C2	250	5.4%

6 結論と今後の課題

6.1 結論

本研究では生徒相互評価の妥当性を，１）構造的要素，２）一般可能性的要素，３）外的要素，４）内容的要素，５）実質的要素，６）影響的要素の６つの側面より検討した。

構造的要素の面は項目応答理論を用いて検証を行い，生徒相互評価が項目応答理論のモデルに適合し一次元性を持つことが部分的に認められ，また高い評価者間信頼性を持つことが支持された。この結果，構造的側面からの生徒相互評価の妥当性は一部で認められたと考える。

一般可能性的要素については，ミスフィットしたクラスがないことや，評価の厳しさに大きな違いがないことからもクラス間あるいは学年間で目立った違いはなく，この側面からの妥当性は認められた。

外的要素の側面は，生徒相互評価と教員評価との間の相関と平均値の比較により検証を行った。クラスにより相関と平均値にややばらつきは認められたものの，高い相関が認められ，平均値の統計的有意差は認められなかった。その結果，外的要素の点からも生徒相互評価は十分な妥当性を持つことが支持された。

内容的要素と実質的要素の側面はアンケートの分析により検証を行い，スピーチと授業の関連性については生徒・教員ともに関連性を認める結果となり，生徒相互評価の適切性について生徒は肯定的にとらえていた。内容的要素，実質的要素のどちらの面からも妥当性が認められたと言える。

さらに，影響的要素については自由記述回答の分析を行い，半数以上の生徒が生徒相互評価についてなんらかのよい影響を受けていると感じていることがわかった。このことから影響的要素についても妥当性が認められたと考える。

これら妥当性の６つの側面の５つの側面で十分な妥当性が認められ，１つの側面で部分的に十分な妥当性が確認された。これらの結果を総合すると生徒相互評価はある程度の妥当性を有すると考えられる。

教育的示唆として以下の３点が挙げられる。第１に，生徒相互評価に一定の妥当性が示されたことで，高校においても教員評価だけでなく生徒相互評価を

評価の一部として取り入れることにより，より多面的で信頼性の高いスピーキングの評価を行うことができると思われる。スピーキングの評価は主観的で（McNamara, 2000）評価が難しいとされ，オーラル・テストの信頼性を確保するため複数の評価者が求められるが（Fulcher, 2003），実際の教育現場では評価者は英語教員1人かALTを含めても2人での評価が限度であろう。生徒相互評価を活用することでより多面的で，妥当性の高いスピーキングの評価につなげることが可能であると考える。

第2に，スピーキングの評価の実用性を高めることに貢献できることである。生徒相互評価を用いると評価を授業時間中に行うことができるので，インタビュー形式などのスピーキングの評価に比べ，評価にかかる労力や時間が少しでも軽減されると思われる。

第3に，アンケートにおいて示されたように，生徒の動機づけを高めたり（Orsmond, et al., 1996），学習目標を明確にする（Luoma, 2004）など，生徒のより積極的な授業参加を促すこともできるであろう。ただし，生徒相互評価の活用にあたっては，平均点が教員評価に比べてやや高くなる傾向にある点や，極端な評価を敬遠する傾向があること（Stefani, 1994）に注意が必要である。スピーキング評価の妥当性や実用性を高めることができれば，授業でより多くのスピーキングの活動を取り入れ，新学習指導要領で求められている4技能を有機的に関連させたコミュニケーション活動をさらに積極的に活用していくことができると考える。

6.2　今後の課題

今後の課題を3点述べたい。

第1に，本研究ではスピーチにおける生徒相互評価の妥当性について検証を行ったが，スピーチは発信のみ（one way）のコミュニケーションであり，スピーキングの評価を行うためにはディスカッションやディベートなど双方向（two way）のタスクに対する生徒相互評価の妥当性の検証が今後の課題の1つである。

第2に，今回の研究ではMessick（1996）の枠組みに基づき，妥当性の6つの要素から検証を行ったが，それぞれの要素についての検証の側面は限定されており，さらにさまざまな観点から調べることが必要であると考える。

第3に，本研究では参加者が項目応答理論の1パラメータ・ロジスティック・モデルに必要とされる100名（大友, 1996）を超えたが，研究者の勤務校でのデータであり，本研究の結果をそのまま一般化するのは難しいと考える。したがってより多様な参加者のデータを基にさらなる研究が求められる。

謝　辞

この研究の機会を与えてくださった（財）日本英語検定協会，選考委員の先生方，特に小池生夫先生に厚く御礼申し上げます。また，常に励ましの言葉をいただいている望月昭彦先生，貴重な助言をいただきました小泉利恵先生，長橋雅俊先生に深く感謝いたします。さらに，この研究に協力していただきました勤務校の先生方にも多大なご助力をいただきました。この場をお借りして，御礼を申し上げます。最後にこの研究を陰で支えてくれた，妻・理香と3人の子供たち・友香，恵理，匠に心より感謝したい。

参考文献（＊は引用文献）

＊秋山朝康.（2000）.『スピーキングテストの分析と評価―項目応答理論を使っての研究―』. *STEP BULLETIN*, vol.12, 67-78.
＊馬場哲夫・片桐一彦・小室俊明・高山芳樹・武井昭江・竹田恒美他.（1997）.『英語スピーキング論―話す力の育成と評価を科学する』. 東京：河源社.
＊Bachman, L.F.（1990）. *Fundamental considerations in language testing*. U.K.: Oxford University Press.
＊Brown, J.D.（1996）. *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
＊Brown, J.D.（Ed.）.（1998）. *New ways of classroom assessment*. Alexandria, VA: Teachers of English to Speakers of Other Languages.
＊Chapelle, C.A.（1999）. Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.
＊Council of Europe.（2001）. *Common European framework of reference for languages: Learning, teaching, assessment*. U.K.: Cambridge University Press.
＊Freeman, M.（1995）. Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education*, 20, 289-299.
＊Fukazawa, M.（2007）. *Validity of peer assessment of*

speaking performance: A case of Japanese high school students. Unpublished master's thesis, University of Tsukuba, Ibaraki, Japan.
＊Fulcher, G.(2003). *Testing second language speaking*. Edinburg Gate, U.K.: Pearson Education.
＊Heaton, J.B.(1975). *Writing English language tests*. Essex, U.K.: London.
＊Hughes, A.(1989). *Testing for language teachers*(2nd ed.). U.K.: Cambridge University Press.
＊Hughes, I.E., & Large, B.J.(1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education*, 18, 379-385.
＊清川英男.(1990).『英語教育研究入門』. 東京：大修館書店.
＊小泉利恵.(2005).『日本人中高生における発表語彙知識の広さと深さの関係』. *STEP BULLETIN*, vol.17, 63-80.
＊Lado, R.（1961). *Language testing: the construction and use of foreign language tests*. London: Longman.
＊Linacre, J.M.(2008). Facets: Rasch-measurement computer program（Version3.64.0）[Computer software]. Chicago: MESA Press.
＊Luoma, S.(2004). *Assessing speaking*. U.K.: Cambridge University Press.
＊Matsuno, S.(2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26, 75-100.
＊松沢伸二.(2002).『英語教師のための新しい評価法』. 東京：大修館書店.
＊McNamara, T.F.(1996). *Measuring second language performance*. London: Longman.
＊McNamara, T.F.(2000). *Language testing*. Oxford: Oxford University Press.
＊Messick, S.(1996). Validity and wash back in language testing. *Language Testing*, 13, 241-256.
＊Miller, L., & Ng, R.(1996). Autonomy in the classroom: Peer assessment. In R. Pemberton, E.S.L. Li, W.F. Or, & H.D. Pierson(Eds.), *Taking control: Autonomy in language learning*(pp.133-146). Hong Kong: Hong Kong University Press.
＊三浦省吾.(2004).『教育データ分析入門』. 東京：大修館書店.
＊文部科学省.(2009).『高等学校学習指導要領』. Retrieved April 25,2009, from http:www.mext.go.jp/a_menu/shotou/new-cs/youryou/kou/kou.pdf
＊Nakamura, Y.（2002). Teacher assessment and peer assessment in practice. *Educational Studies*, 44, 203-215.
＊長橋雅俊.(2008).『コーパス分析とラッシュ・モデルを用いたライティング・テストでの困難度比較』. *STEP BULLETIN*, vol.20, 95-115.
＊大友賢二.(1996).『項目応答理論入門』. 東京：大修館書店.
＊大友賢二・中村洋一.(2002).『テストで言語能力は測れるか～言語テストデータ分析入門～』. 東京：桐原書店.
＊Orsmond, P., Merry, S., & Reiling, K.(1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21, 239-250.
＊靜哲人・竹内理・吉澤清美.(2002).『外国語教育リサーチとテスティングの基礎概念』. 大阪：関西大学出版部.
＊Stansfield, C.W., & Kenyon, D.M.（1995). Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. In A. Cumming & R. Berwick（Eds.), *Validation in language testing*. (pp.124-153). Clevedon, U.K.: Multilingual matters.
＊Statistical Package for the Social Sciences(SPSS).(2008). SPSS(Version 17.0)[Computer software]. Chicago: SPSS Inc.
＊Stefani, L.A.J.(1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, 19, 69-75.

資　料

資料1：評価基準表

項目	基準	Basic（初級）		Intermediate（中級）		Advanced（上級）	
		A1	A2	B1	B2	C1	C2
発音	声の大きさ	**聞き取るのが難しい。** 声が小さかったり，発音が不明瞭で聞きとることが困難である。		**聞き取ることが可能である。** 声の大きさはまずまずで，発音が不明瞭な点もあるが，聞きとることは可能である。		**良く聞き取ることができる。** 声の大きさもよく，発音も明瞭で良く聞きとることが出来る。	
	発音	聞き取りが難しい 聞き取り：×	部分的に聞きとれる 聞き取り：△	何とか聞きとれる 聞き取り：△＋	大体聞きとれる 聞き取り：○	＋分聞きとれる 聞き取り：○＋	大変良く聞きとれる 聞き取り：◎
正確さ	文法	A1	A2	B1	B2	C1	C2
		間違いが多い。 理解を妨げるような文法的間違いが多く，簡単な文法に限られている。		**間違いはあるが，まずまず正確** 理解を妨げるような文法的間違いが少しある。		**間違いがあまりない。** 文法的誤りが少なく，理解を妨げることはない。高度な文法も用いている。	
		理解困難なほど間違いが多い 文法：×	部分的理解に止まるほど間違いが多い 文法：△	理解を妨げる間違いが数カ所 文法：△＋	理解を妨げる間違いが少し 文法：○	間違いが少なく，＋分理解できる 文法：○＋	間違いがほとんど無くよく理解できる 文法：◎
流暢さ	とぎれのなさ 強弱のリズム	A1	A2	B1	B2	C1	C2
		話が中断してしまう。 発話がしばしば中断したり，強弱によるリズムがあまりない。		**時々，話がとぎれることがある。** 時々とぎれることがあるが，強弱によるリズムは少しある。		**話が自然に流れる。** 発話が自然であり，強弱によるリズムがある。	
		長い沈黙あり 流れ：×	細切れの発話 流れ：△	とぎれがち 流れ：△＋	時々とぎれる 流れ：○	ほとんど自然 流れ：○＋	自然 流れ：◎
態度	ｱｲｺﾝﾀｸﾄ ｼﾞｪｽﾁｬｰ 積極性	A1	A2	B1	B2	C1	C2
		積極的に話そうする姿勢が弱い。 アイ・コンタクト，ジェスチャー，表情等があまりない。		**積極的に話そうとしている。** アイ・コンタクト，ジェスチャー，表情等を時々使っている。		**積極的・効果的に話している。** アイ・コンタクト，ジェスチャー，表情等が多く，効果的に使われている。	
		積極的ではない 態度：×	あまり積極的でない 態度：△	少し積極的 態度：△＋	おおむね積極的 態度：○	積極的 態度：○＋	非常に積極的 態度：◎

総合評価	内容も含めた総合評価	A1	A2	B1	B2	C1	C2
		メッセージが十分に伝わらない。 内容の理解が難しく，発表として不十分な点が多い。		**メッセージが大体伝わる。** 一部わかりづらい点はあるが，主旨は理解することができる。まずまずの発表。		**メッセージがよく伝わる。** わかりやすく，内容を良く理解することができる。魅力的な発表。	
		伝わらない 発表：×	部分的に伝わる 発表：△	何とか伝わる 発表：△＋	おおむね伝わる 発表：○	＋分伝わる 発表：○＋	効果的に伝わる 発表：◎

資料 2：アンケート項目（例）

〈2 ～ 4 時間目の生徒による相互評価について〉

(12) 今回のスピーチは，これまで 1 年生の英語の授業で行ってきたスピーキングの活動と関連していると思う。
①とてもそう思う　②ややそう思う
③あまりそう思わない　④全くそう思わない
⑤わからない

(13) スピーチの評価項目（発音・正確さ・流暢さ・態度・総合）は，これまで 1 年生の英語の授業で行ってきたスピーキングの活動と関連していると思う。
①とてもそう思う　②ややそう思う
③あまりそう思わない　④全くそう思わない
⑤わからない

(18) 相互評価は，発表者のスピーキングの能力を評価していると思う。
①とてもそう思う　②ややそう思う
③あまりそう思わない　④全くそう思わない
⑤わからない

(30) 相互評価をすることで，英語学習や他の面で勉強になった点があれば記入してください。

（　　）

資料 3：生徒相互評価と教員評価の記述統計

	クラス	A	B	C	D
	発表者数	36	35	35	39
生徒相互評価	*M*	3.97	3.91	3.79	3.94
	SD	.77	.79	.78	.77
	n	30	28	30	29
教員評価	*M*	3.78	3.51	3.59	3.92
	SD	.93	.94	.80	.81
	n	3	3	4	4

（注）*M* = 平均値；*SD* = 標準偏差；*n* = 評価者の人数

資料 4：アンケートの回答（例）

【肯定的意見】

・自分はたぶん相互評価がなかったら，集中して人のスピーチを聞かないと思う。だから，相互評価があって良かった。集中して聞けた。
・きちんと聞こうという気になり，リスニングする集中力が続くようになった。
・スピーチをする上で気をつけるポイントを改めて認識することができた。
・みんながどれくらいスピーチの能力があるのかわかったので，これからの自分もそれを基準として考え，もっと上達した実用的な英語能力を目指そうとかんばる。
・人のスピーチを評価することで，良い点や悪い点などを自分に重ねて冷静に見ることができるから，いい勉強になった。
・人のスピーチを聞いて「あ～，いいことだな～」だと思って，やってみたい点がある。
・自分のレベルを理解できて，目標にしようと思う人物を見つけられた。今後その目標に向かって頑張れる。
・自分と他人を比べて，自分に足りないところや自信を持てるところが見つかった。
・評価の練習により，基準がよくわからなかったのでより適切な評価ができるようになると思う。
・友達の評価をして，自分の不十分な点や補う点がわかったこと。
・人前で英語を話す機会はめったにないし，英語によるいろいろな人の意見が聞けて非常にためになった。

【中立的意見】
・人前にたってスピーチすることの大変さ。
・全体的にそこそこできていれば高い評価を得られる。

【否定的意見】
・客観性を方法的に高められること，しかしあまり長続きはしない可能性が高いことがわかった。
・自分と相手を比べることの難しさを感じた。

（注）記述回答の表現は，生徒の回答のままである。