

## 教師の評価コメントに基づいたルーブリック作成の取り組み

安高紀子・品川なぎさ

### 要旨

本稿は、聖心女子大学の短期留学生日本語クラスで実施している調査発表に対する評価について、教師間で評価の観点や基準を共有し、評価の信頼性を高めること、また、授業改善に活かすことを目的として行ったルーブリック作成の取り組みについて報告するものである。まず、調査発表に対する教師の評価コメントを分析した結果、調査の内容面へのコメントが多いこと、教師によって注目している観点到に相違があることがわかった。次に、評価コメントの分析に基づき、ルーブリックを作成し、それをを用いて教師2名で評価を行った。そして、評価結果について検討し、ルーブリックの修正を行い、再評価した。その結果、ルーブリックの修正により、一致度は概ね上昇し、評価者間の差を低減できた可能性が示唆された。ルーブリックを作成し、評価する過程を通じて、教師は何をどう評価するかを改めて考え、教師間で評価の観点を共有できたことが評価の信頼性を高めることにつながったと思われる。

### キーワード

調査発表、ルーブリック、評価コメント、評価の観点

## 1. はじめに

### 1.1 背景

聖心女子大学の短期留学生日本語クラス<sup>(1)</sup>では、授業の一環として、日本人を対象としたインタビュー調査やアンケート調査を実施しており、授業で学生が行った調査発表に対しては、教師が評価を行っている。本クラスはティームティーチングで3名の教師が担当しており、評価を担当する教師は毎回異なる。調査発表の評価は、総合的に評価する包括的評価が行われてきたが、評価の観点や基準はあまり明確にされていなかった。そのため、例えば同じ「A」評価でも、教師によって、どういう点についてどのように評価しているかにばらつきがあり、評価の信頼性に問題が生じていた可能性がある。また、本クラスでは、調査発表は全3回実施され、発表のスキルを身に付けることを目指し、形成的評価が行われていた。しかし、これまでの評価では、学生の達成度を具体的に把握することが難しく、授業の改善に活かすことが十分にできていなかった。

### 1.2 目的

本研究が対象としている調査発表や、スピーチや作文などのパフォーマンスに対する評価では、総合的に評価する包括的評価、いくつかの観点到に分けて評価する分析的評価などが行われている。包括的評価には学習者のパフォーマンスを総合的に評価でき、評価にかかる時間も少ないというメリットがあるが、一方で、パフォーマンスの長所や短所が明らかにできない(近藤ブラウン 2012)という問題点もある。また、マクナマラ(2004)

は、パフォーマンスに対して、全体的に捉える包括的評価よりも、様々な面について個別に評価すべきだとし、そのためには個別に評価尺度を設定する必要性を指摘している。

パフォーマンスに対する評価の方法としては、ルーブリックを用いて質的に評価する方法が注目されている。ルーブリックとは、「成功の度合いを示す数値的な尺度とそれぞれの尺度に見られる認識や行為の特徴を示した記述語からなる評価指標」(石井 2010)で、評価の信頼性の保証、教師間の評価基準の共有に有効だとされている。

そこで、これまで本クラスで行われてきた調査発表の評価における問題点を改善する方法として、調査発表で評価すべき観点を定め、観点別に到達段階を記述したルーブリックを作成し、それを用いて評価を行う取り組みを行った。ルーブリックの作成、それを用いた評価という一連の過程を通して、教師間で共通した評価の観点や基準を持ち、評価の信頼性を高めることを目的とする。

## 2. ルーブリックの作成方法

本研究では、次のようなステップで本クラスの調査発表のためのルーブリックを作成する。まず、これまで教師がどのような観点で調査発表の評価を行っていたのかについて、教師が授業報告に記した評価コメントを分析し、評価の観点の洗い出しを行う。次に、抽出された観点に基づき、ルーブリックを作成する。ルーブリックの作成手順は、黒上

(2016)を参考に、課題の到達目標を設定して、それに対する達成度を評価項目ごとに段階的に記述したルーブリックを作成する。そして、作成したルーブリックを用いて、実際に評価を行う。さらに、レベルごとのパフォーマンスにどのような特徴があったのかを記述する西岡(2015)、田中(2011)の方法に倣い、調査発表で見られた特徴をルーブリックの記述文に反映させ、修正を行う。

## 3. 教師の評価コメントの分析

### 3.1 分析データと分析方法

分析したデータは、2013年度前期～2016年度前期の間に行われた全14回の調査発表<sup>(2)</sup>(アンケート調査7回、インタビュー調査7回)に対して、教師3名が記述した評価コメント242件である。分析方法は、教師が記した評価コメントの文言について、何について評価したものか、次のⅠ～Ⅲの判定基準に基づいて判定を行い、タグ付けした。タグ付けは、本クラスで評価を担当した2名の教師が個別に行い、相違のあったものについては、協議し、両者の合意のもと、最終的な判定を決定した。

- Ⅰ. コメントに評価の観点である言葉が明示されている。
- Ⅱ. コメントに評価の観点と思われる言葉や表現が示されている。
- Ⅲ. コメントに何に対する評価か、観点が明示されていない。

例えば、「目的が明瞭」(775<sup>(3)</sup>)というコメントは、調査の「目的」についての評価であり、何についての評価なのかが明確に示されており、判定基準Ⅰにあたりと判断した。「知りたいことが明瞭」(49)は、「知りたいこと」というのは、調査を行う「目的」に当たる言葉であり、これは判定基準Ⅱに該当すると判断した。また、「全体的によ

い」(I112)は、総合的な評価であり、評価の観点を示されておらず、判定基準Ⅲに該当するとした。上記のⅠ～Ⅲの基準によって、コメントから判定された観点とコメントの例を表1に示す。

表1 教師の評価のコメントと観点

	コメント No.	教師のコメント	観点
Ⅰ	ア75	<u>目的</u> が明確。	目的
	イ119	<u>構成</u> がよく、わかりやすい。	構成
	ア87	<u>全体的にまとまっている</u> 。	まとまり
	ア89	<u>説明</u> が丁寧にできていた。	説明
	ア41	<u>分析</u> がよくできている。	分析
	イ43	<u>時間</u> オーバー。	発表時間
	ア91	グラフを用いて、視覚的な <u>工夫</u> あり。	発表の工夫
Ⅱ	イ49	<u>知りたいこと</u> が明瞭。	目的
	ア21	<u>流れ</u> がわかりやすい。	構成
	ア27	<u>語彙の説明</u> など、聞き手への配慮あり。	説明の工夫
	イ106	<u>イントネーションの問題</u> で聞きとりにくいところあり。	発音
	ア93	<u>結果に基づいた意見</u> ではなく、自分の考えを述べている。	考察
	イ102	<u>レジュメ</u> はわかりやすくまとめられていた。	準備
	イ24	原稿を見ないで発表。	発表
Ⅲ	イ112	全体的によい。	不明

評価コメントに観点をタグを付けた結果、全部で16の観点(目的、目的と質問の一貫性、テーマ選択の観点、構成、まとまり、説明、説明の工夫、分析、考察、発音、文法・語彙、発表時間、発表準備、発表の工夫、発表、不明)が抽出された。これらのうち、不明を除く15の観点について、観点として関連性の高いものをまとめ、「目的」「構成」「説明」「分析・考察」「発音」「文法・語彙」「プレゼンテーション」の7つの評価の観点到分類した。表2に調査発表に対する評価の観点到まとめる。

表2 調査発表の観点

評価の観点	コメントに見られた観点
目的	目的、目的と質問の一貫性、テーマ選択の観点
説明	説明、説明の工夫(言い換えなど)
分析・考察	分析、考察
構成	構成、まとまり
発音	発音(アクセント、イントネーション、清濁など)
文法・語彙	文法・語彙
プレゼンテーション	発表時間、発表(アイコンタクトなど)、 発表準備(レジュメ)、発表の工夫(グラフの使用)

### 3.2 評価コメントの分析結果と考察

#### 3.2.1 観点別コメントの割合

調査発表の評価コメント全 242 件について、観点ごとの割合を算出した結果を図 1 に示す。図 1 のグラフの通り、もっともコメントの多かった観点は「分析・考察」で 22%、次いで「構成」17%、「目的」16%という順であった。全コメントの内、「目的」「説明」「分析・考察」の 3 つが約半数を占めており、これらの観点は調査の内容に関するものである。つまり、調査発表に対する教師の評価コメントは、調査の内容に関するものが多く、調査発表の評価において、内容面に注目した評価が行われていることがわかった。

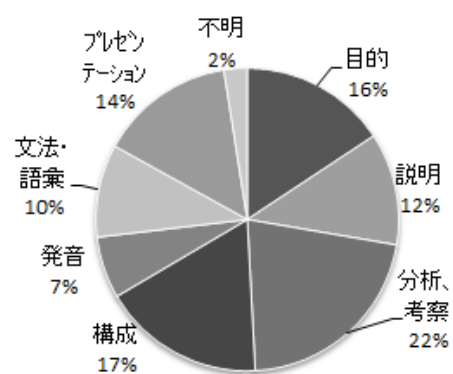


図 1 観点別コメントの割合

#### 3.2.2 各教師の評価の観点

次に、評価を担当した 3 名の教師ごとに、コメント数が多かった上位 3 つの観点を比較した結果 (表 3)、教師 3 名の観点別のコメントの割合やその順位には異なる傾向がみられた。教師 A は「目的」「分析・考察」「説明」という調査の内容に関するコメントが 6 割を占め、評価において内容面を重視している。教師 B は発表の際の「プレゼンテーション」や「発音」「文法・語彙」が多く、プレゼンテーション技術や日本語能力に注目して評価を行っている。また、教師 C は発表の内容に関わる「分析・考察」が約 3 割を占め、他には「構成」という全体の流れや日本語能力に関する「文法・語彙」と、多様な観点から評価をしていることがわかる。

本クラスの調査発表に対しては、これまで教師自身が持つ評価の観点に基づいた包括的評価が行われてきたが、各教師が評価で注目していた観点には異なる傾向が見られた。これは、日本語小論文の評価について、評価プロセスや決定要因を分析した田中・坪根 (2011) の「どの要素を優先させて評価するかについては日本語教師の間に共通の認識が認められなかった」という結果と同様のものである。このように、本クラスの調査発表の評価でも、教師によって評価している観点には相違が認められた。

表 3 教師別コメント数の多かった観点上位 3 つ

教師 A	教師 B	教師 C
目的 22%	プレゼンテーション 22%	分析・考察 29%
分析・考察 22%	発音 18%	構成 18%
説明 17%	文法・語彙 14%	文法・語彙 14%

### 4. ルーブリックの作成と評価

黒上 (2016) を参照し、評価コメントの分析によって抽出された 7 つの観点を評価項目として、評価項目ごとに到達目標を設定し、到達目標に対する達成度を 4 段階

(A. B. C. F) で記述したルーブリックを作成した(資料1)。ルーブリックの作成は、評価対象となる留学生は含まず、調査発表の評価を担当する教師のみで作成を行った。

#### 4.1 ルーブリックによる評価と一緻度

2016年度後期に行われたアンケート調査の発表<sup>(4)</sup>をビデオに録画したものを、本クラスを担当する教師2名が視聴し、作成したルーブリックを用いて、評価を行った。評価の結果について、評価者間の評価の一緻度をカップ係数により算出した。その結果、全体としては、全49項目中27項目が一致し、一緻度は $\kappa = .55$ で、中程度の一緻であった。評価項目別にみると、「目的」「分析・考察」「文法・語彙」はかなり高い一緻だったが、「構成」「プレゼンテーション」は中程度、「説明」「発音」はやや低い一緻であった。このように、評価項目によって一緻度に差があることがわかった。

#### 4.2 ルーブリックの修正

ルーブリックを用いて評価を行った後、教師間で評価の結果を照らし合わせ、評価基準の再検討を行った。評定が一致しなかったものについては、教師それぞれがどのように評価したのかを話し合い、評定に差異が生じた要因を確認した。また、西岡(2015)、田中(2011)のルーブリック作成方法を参照し、留学生の調査発表にみられた特徴を具体的に挙げ、ルーブリックに記述すべき点を話し合い、ルーブリックの再検討を行った。

再検討した内容をふまえ、ルーブリックの修正を行った(資料2)。修正した点は①記述文の詳細化と、②評価尺度の追加の2点である。まず、①記述文の詳細化については、評価が分かれた要因として、記述文に曖昧さがあったため、どの尺度に位置付けるべきか、判断に迷ったものがあり、段階ごとの記述文をより詳細にした。例えば、評定の一緻度がやや低かった「発音」の記述文に関しては、量と質の2つの側面に関する記述を追加した。発音のミスが多いか少ないという量的な面、発表は聞きとりやすいかどうかという質的な面に関して、尺度ごとに段階的な記述をして調整を行った。「説明」についても、伝えている情報の量が適切か不足しているか、情報量という面での記述を追加した。また、「構成」については、留学生が行った発表で観察された、発表内容の重複などの問題点を具体的に挙げ、それらの点を記述文に反映させた。このように、記述をより明確にすることで精緻化を図った。

次に、②評価尺度の追加については、4段階の最上位「A」の上に、さらに「S」という尺度を追加した。評定で同じ「A」とされた発表でも、かなりの幅があり、その中には特に評価すべき優れた特徴がみられるものがあった。そのため、その差異を明らかにする必要性から、「A」の上にもう1つ「S」という尺度を追加することとした。「S」を追加したのは、すべての評価項目ではなく、特に評価すべき特徴がみられた「目的」「説明」「分析・考察」「発音」「文法・語彙」の5つの項目であった。

#### 4.3 修正したルーブリックによる再評価と一緻度

修正したルーブリックを用いて、1回目と同じ評価者が、調査発表のビデオを再度視聴し、評価を行った。ルーブリックの修正として、上述の通り、①記述文の詳細化と、②評価尺度の追加の2点について修正が行われており、評定結果にはこの両方の点が影響し

ていると考えられる。そこで、これらの修正がどのように影響しているか、それぞれの修正点ごとに検証を行った。まず、①記述文の詳細化について検証すると、評定の一致度は  $\kappa = .75$  でかなり高い一致であった。次に、②4段階から5段階に尺度を追加した点については、一致度は  $\kappa = .70$  で、こちらもかなり高い一致であった。いずれの結果も、修正前の  $\kappa = .55$  よりも高く、評価者間の一致度は上昇したことが認められる。以上の結果から、ルーブリックの修正によって、評定のばらつきを低減できた可能性が示唆された。

#### 4.4 評価項目別の評定

評定の一致度について、ルーブリックの修正前と修正後を比較したものを表4に示す。

表4 ルーブリック修正前と修正後の評定の一致度

一致度	ルーブリック修正前	ルーブリック修正後
かなり高	「目的」「文法・語彙」 「分析・考察」	「目的」「文法・語彙」 「発音」「構成」「プレゼンテーション」
中	「構成」「プレゼンテーション」	「分析・考察」「説明」
やや低	「発音」「説明」	—

表4のとおり、修正前と修正後の一致度を比較すると、「目的」「文法・語彙」の一致度はほぼ変わらず、かなり高い一致という結果であった。「構成」「プレゼンテーション」は、修正前は中程度の一致だったが、修正後はかなり高い一致となった。そして、「発音」は、修正前は  $\kappa = .23$  であったが、ルーブリックの記述文の修正で、量と質の2つの側面から詳細に記述したことで、 $\kappa = .82$  となり、評価者間ではほぼ一致する結果となり、一致度は大きく上昇した。「説明」では、修正で情報量に関する記述を追加したことで、修正前の  $\kappa = .21$  から、修正後は  $\kappa = .48$  で、中程度の一致という結果となり、一致度の上昇が見られたが、全項目の中ではもっとも一致度が低かった。「説明」の一致度が低かった要因としては、次のことが考えられる。ルーブリックを用いた評価後に行われた教師間での検討において、「説明」の評価でわかりにくいと感じる場合、「日本語能力の問題なのか、説明の内容自体の問題なのか、その判断が難しい」という意見があった。つまり、他の評価項目の要素が関係している恐れがある。その点が影響した可能性が考えられる。また、「分析・考察」は、修正前は  $\kappa = .78$  でかなり高い一致であったが、修正後は  $\kappa = 0.53$  で中程度の一致となり、一致度が下がった。この「分析・考察」は、「根拠の有無」「意見は説得力があるか」という点で判断される。だが、「根拠の有無」は、根拠が示されているかどうかだけでなく、示された根拠が根拠として妥当なものであるのか、また、「説得力のある意見かどうか」は、説得力のある意見とはどういうものか、教師がそれらをどのように判断するかによる。したがって、教師の主観が影響していることが要因ではないかと推測される。

全評価項目の中で、一致度がもっとも低かったのは「説明」、2番目は「分析・考察」で、これらの項目は、調査の内容面の評価に関する項目で、他の項目と比べ、評価にずれが生じやすいという傾向がうかがえる。「発音」や「文法・語彙」のような日本語能力に

関するものは、評価の対象が発表でも作文でも、日本語教師であれば、ある程度共通したものさしで測ることが可能だと思われる。しかし、調査発表の内容面に関しては、どのように評価するかは教師によって違いが出やすいということではないだろうか。3.2.1.の教師の評価コメントの分析結果の通り、内容面に関するコメントは全評価コメントの約半分を占めており、評価において注目される観点ではあるが、他の観点に比べ、教師によって評価に差が生まれやすい傾向がある。評価の信頼性を高めるためには、教師間で共通した評価基準を持つことが重要であろう。

## 5. おわりに

本研究では、短期留学生日本語クラスで実施している調査発表に対して、教師の評価コメントを分析し、教師がどのような点を評価していたのかを明らかにした。次に、分析によって抽出された観点を基に、ルーブリックを作成し、そのルーブリックを用いて、教師2名で調査発表の評定を行った。そして、評定結果を基に、ルーブリックの検討、修正を行い、再度、評価した。その結果、評価者間の評定の一致率は概ね上昇し、評価で生じていた差を低減できた可能性が示唆された。

本研究で行った評価コメントの分析による観点の抽出、ルーブリックの作成、ルーブリックを用いた評価、記述文の検討、修正といった一連の取り組みを通して、教師自身の評価の傾向を知るとともに、教師間で評価の観点や評価基準を共有することができた。また、その過程で得た経験が評価の信頼性を高めることにつながったと思われる。

今後の課題としては、今回作成したルーブリックを用いて、新たに評価を行い、発表に見られた特徴を記述に追加するなどの修正を加え、ルーブリックをさらにより良いものにしていくことである。また、現時点では、本ルーブリックは評価される側の学生と共有できておらず、今後は、学生へのフィードバックにも活用し、評価を可視化することによって、学生の調査発表のスキルの向上につなげていきたい。

(安高紀子あたかのりこ・聖心女子大学・nrkataka@gmail.com)

(品川なぎさしながわなぎさ・聖心女子大学・shinagawa618@gmail.com)

## 謝辞

調査にご協力くださった聖心女子大学の短期留学生のみなさん、また、小川早百合先生、関かおる先生、高柳なな枝先生には、調査に関する貴重なご助言を賜り、心より感謝申し上げます。

## 注

1. 本クラスは協定校からの短期留学生対象のクラスで、日本語レベルは中級～上級前半程度。授業期間はおよそ4か月(90×45回)。使用テキストは『トピックによる日本語総合演習-テーマ探しから発表へ-上級』(スリーエーネットワーク)である。
2. 全14回の調査発表で各教師が評価をした回数は、教師Aが6回、教師Bが4回、教師Cが4回であった。
3. アはアンケート調査、イはインタビュー調査の略、また、数字はコメントナンバーを示す。

調査報告

安高紀子・品川なぎさ/アカデミック・ジャパニーズ・ジャーナル9(2017)28-36

4. アンケート調査のトピックは「宗教」で、7名の留学生が発表を行った。発表時間は一人7分であった。

資料1 調査発表のルーブリック (修正前)

評価項目	A(優)	B(良)	C(可)	F(不可)
目的と質問項目	調査目的が明確で、それに関する適切な質問項目が設定されている。	調査目的は示されているが、質問項目との関連が弱い。	調査目的は示されているが、質問項目とずれがある。	調査目的がわからない。目的と質問項目との関連がない。
構成	構成ができており、流れがよい	構成ができています。	構成に問題がある	構成されていない
説明	調査内容、結果について、端的に説明ができています	調査内容、結果について、適切に説明ができています	調査内容、結果について、冗長、または説明不足のため、理解しにくい。	調査内容、結果について、冗長、または説明不足のため、理解できない。
分析・考察	根拠を示し、説得力のある意見を述べている。	根拠を示しているが、意見は、ごく一般的で、ありふれている。	根拠を示しているが、解釈に誤りがある、または、客観性がない	根拠がなく、感想を述べている。
発音 (アクセント、イントネーション、清濁など)	発音が正確、かつ明瞭で、聞きとりやすい。	発音のミスはあるが、聞きとれる	発音のミスが多く、聞きとりにくい	発音が不正確で聞き取れない
文法・語彙	文法・語彙のミスがほとんどなく、適切に使用できている。	文法・語彙のミスはあるが、理解に負担がない	文法・語彙ミスのため、理解に負担を感じる	文法・語彙ミスのため、理解できない。
プレゼンテーション ①アイコンタクト ②発表の工夫 (ことばの定義や言い換え、グラフの使用など) ③発表時間の厳守	(左記のうち)3つできている。	(左記のうち)2つできている。	(左記のうち)1つできている。	できていない

資料2 調査発表のルーブリック (修正後)

評価項目	S	A	B	C	F
目的	調査の目的や理由が明確で、調査する意義が感じられる。調査の観点もよい。また、目的に合った適切な質問が設定されている。	調査の目的や理由が示されており、調査のテーマは、自分の知識や経験に基づき、設定されている。また、目的に合った質問が設定されている。	調査の目的や理由は示されているが、調査のテーマはごく一般的、または抽象的。また、調査のテーマと質問との関連が若干弱い。	調査の目的、または理由がはっきりしていない。また、調査のテーマ設定が明確でないため、質問との間にズレがある。	調査の目的、または理由が示されていない。
構成		発表の構成(「調査目的」「質問と結果」「考察」)に沿って、発表ができています。	発表の構成に沿ってはいませんが、内容が重複したり、対応していない箇所がところどころある。	発表の構成と発表内容が対応していない部分があるため、わかりにくい。	発表が構成されておらず、言いたいことを順に述べている。
説明	発表内容について、聞き手への配慮(ことばの定義や言い換え)をしながら、わかりやすく、かつ端的に説明ができています。	発表内容について、聞き手が理解するのに適切な情報量で説明ができています。	発表内容について、部分的に説明が足りない、または、まとも次に欠ける部分があるが、情報を伝えることはできています。	発表内容について、説明不足、または、冗長なため、情報量の過不足があり、言いたいことを理解するのに負担を感じる。	発表内容について、説明不足、または、冗長で、言いたいことが伝わらない。
分析・考察	調査結果に基づき、根拠を示しながら、説得力のある意見を述べている。	調査結果に基づき、根拠を示しながら、自分の意見を述べている。	調査結果に基づき、自分の意見を述べているが、ごく一般的で、ありふれたものである。	調査結果に基づき、意見を述べているが、客観性に欠けている。または、結果の解釈に誤りがある。	調査結果に基づいた意見ではなく、自分の感想を述べている。
発音 (アクセント、イントネーション、清濁など)	発音が正確、かつ明瞭である。	発音のミスが少なく、発音は聞きとりやすい。	発音のミスがみられ、聞き取りにくいこともあるが、だいたい理解できる。	発音のミスが多く、聞きとりにくい。発音ミスのため、理解しにくいことがある。	発音が不正確で聞き取れない。
文法・語彙	文法・語彙の使用が適切で、バリエーションも豊富。	文法・語彙のミスがほとんどなく、適切に使用できている。	文法・語彙のミスはあるが、理解に負担がない	文法・語彙ミスのため、理解に負担を感じる	文法・語彙ミスのため、理解できない。
プレゼンテーション ①アイコンタクト、姿勢 ②発表準備(レジメなど) ③発表時間の厳守		(左記のうち)3つできている。	(左記のうち)2つできている。	(左記のうち)1つできている。	できていない

参考文献

石井英真 (2010) 「VI 教育目標と教育評価の関係 7 ルーブリック」, 田中耕治 (編) 『よくわかる教育評価』 ミネルヴァ書房, 48-49.



黒上晴夫 (2016) 「パフォーマンス評価としてのルーブリック」, 第 40 回 AJG 研究会「アカデミック・ジャパニーズにおけるパフォーマンス評価としてのルーブリックを考える」2016. 11. 05 配付資料

近藤ブラウン妃美 (2012) 『日本語教師のための評価入門』くろしお出版

田中耕司 (2011) 『パフォーマンス評価－思考力・判断力・表現力を育む授業づくり』ぎょうせい

田中真理・坪根由香里 (2011) 「第二言語としての日本語小論文における good writing 評価－そのプロセスと決定要因－」『社会言語科学』14(1), 210-222.

西岡加名恵 (2015) 『新しい教育評価入門－人を育てる評価のために』有斐閣コンパクト

ティム・マクナマラ (著) 伊東祐郎・三枝令子・島田めぐみ・野口裕之 (監訳) (2004) 『言語テスト概論』スリーエーネットワーク (McNAMARA, T. (2000) *Oxford Introductions to Language Study :Language Testing*, Oxford: Oxford University Press)