大学スペイン語教育での口頭コミュニケーション能力評価における評者間信頼性について

Inter-Rater Reliability in Oral Communication Ability Assessment in a College Level Spanish Language Program

志柿　光浩（東北大学）
三宅　禎子（岩手県立大学高等教育推進センター）

Abstract

The authors have conducted research with the support of grants from the Japan Society for the Promotion of Science under the project titled "Development of a Spanish Speaking Ability Assessment Method Suitable for University Foreign Language Education Program"（Research Period: Fiscal Years 2018-2023）. The aim of the study was to clarify the necessary settings to achieve validity, reliability, and beneficial backwash while ensuring the feasibility of conducting operational assessments in the areas of Spoken Interaction and Spoken Production within the Spanish language education curriculum offered as part of the university's general education program, even in challenging educational environments. With the grants, native speaker teaching assistants were hired. Combined with the university's Teaching Assistant program, two assistants were assigned to each class session to provide support for the instruction and evaluation of oral communication abilities. The primary focus of this paper is to present the results of inter-rater reliability tests conducted on various performance assessments carried out by multiple native speaker teaching assistants. Additionally, the paper discusses the significance of employing native speaker teaching assistants in foreign language education and emphasizes the importance of involving multiple raters in performance evaluations.

キーワード：スペイン語教育、口頭コミュニケーション、パフォーマンス評価、評者間信頼性、テストの実行可能性

１．はじめに

筆者らは「大学外国語教育プログラム内評価に適合したスペイン語スピーキング能力測定手法の開発」という課題で、日本学術振興会科学研究費補助金の交付を受け、実践研究を行ってきた（研究期間：2018 年度～ 2023 年度）。研究の目的は、大学一般教育課程の一環として実施されているスペイン語教育カリキュラムにおいて口頭でのやりとり及び口頭での産出の領域における運用能力評価を実施していくにあたって、妥当性（validity）・信頼性（reliability）・有益な波及効果（beneficial backwash）を確保しつつ、現場の厳しい教育環境の中でも、恒常的にそのような評価が可能となるレベルの実行可能性（feasibility）を実現するには、どのような設定が必要かを明らかにすることであった。

上記研究目的を設定した背景には、日本の大学における外国語教育環境へのパフォー

マンス評価の導入は、実行可能性の点で様々な障害を伴うという現実がある。研究活動に忙しい院生ティーチング・アシスタントやそれぞれの事情で忙しい学外からの授業補助者を集めて評価方法の擦り合わせをしたりすることは現実には非常に難しい。また、評価のための詳細なルーブリックを用意しても、授業内の短時間で利用するのは現実的ではない。また、学期終了期日から成績報告締め切りの期日までが短く、本稿で報告するようなパフォーマンス課題の評価を短期間に済ませることは、評価者には大きな負担である。そのような条件下で、評価者の事前訓練や評価に必要な最低限の信頼性を確保できるか否かという点が、今回の実践研究における重要な検証課題であった。言い換えるならば、教育活動の負担が過重にならない設定で行うパフォーマンス評価で、一定の信頼性を確保することはできるか、実践の中で明らかにしようとしたわけである。

具体的には、2018 年度から 2021 年度まで日本のある大学の一般教育課程スペイン語科目の授業において母語話者授業補助者を採用し、大学のティーチング・アシスタント制度と併せ、毎回の授業に２名の母語話者授業補助者を置き、口頭コミュニケーション能力の指導と評価に携わってもらった。

本稿では、複数の母語話者授業補助者に行ってもらった数種のパフォーマンス評価の評点についての評者間信頼性検定の結果を中心に報告し、併せて外国語教育における母語話者授業補助者の必要性とパフォーマンス評価における複数評価者の必要性について得られた示唆について報告する。

なお、本稿が報告する４年間の実践過程のうち、後半の２年間は Covid-19 の世界的感染爆発の時期と重なってしまった。2020 年度は一年を通して全ての授業がオンラインで行われた。オンラインで母語話者とのやりとり活動中心の教室活動設計を行なったが、それまでに一定程度確立していた指導ルーティンの大幅な見直しを迫られた。母語話者による評価活動にも大きな制約が生じ、細かな点での意思疎通を図るのが難しかった。2021 年度は教室での対面授業に戻すことができたが、座席間隔を空ける、なるべく近くに寄らないなど、以前のようなやりとり活動を行うには制約が残った。このような中で、母語話者授業補助者による評価活動にも制約が生じ、感染の影響が出る前の最初の２年間とは大きく状況が変わる中での実践活動となった。この点については、各年度の評価活動についての報告の中で少し詳しく触れることにする。

２．口頭コミュニケーション能力の位置づけ

当該大学のスペイン語科目は初年度実施の科目については CEFR（*Common European Framework of Reference for Language Learning*, Council of Europe, 2001 & 2020）の A1 レベル、２年目実施の科目については同 A2 レベルを到達目標と規定していた。筆者らが担当した期間、２年目実施科目の履修者からスペイン語の国際標準試験 DELE（Diploma de Español como Lengua Extranjera, [Diploma of Spanish as a Foreign Language]）の A2 レベルの合格者が複数出ていたことから、このような目標設定は妥当であったと考えられる。

従来より筆者らは、このような到達目標を実現するには、特に口頭コミュニケーション能力の獲得を重視した指導を行うべきであるという立場で授業を続けてきており、本稿が報告する期間においても同様であった。

ところで、CEFR は外国語運用能力を次ページ表１に示されたような領域に分類している。

表１．CEFR が提示する言語活動の分類（Council of Europe, 2020, 33）

	創造的・対人的 言語使用 Creative, Interpersonal Language Use	目的遂行的 言語使用 Transactional Language Use	評価的・課題解決的 言語使用 Evaluative, Problem- Solving Language Use
受容活動 RECEPTION	例）余暇活動として読書する。	例）情報や主張を知るために文章を読む。	（前項「情報や主張を知るために文章を読む」と共通）
産出活動 PRODUCTION	例）経験したことを一定時間一人で語る。	例）情報を伝えるために一定時間一人で語る。	例）事例紹介のために一定時間一人で語る。
やりとり活動 INTERACTION	例）会話をする。	例）モノやサービスを得る。情報を交換する。	例）討論を行う。
仲介活動 MEDIATION	例）コミュニケーションの橋渡しをする。	例）テクストの意味を橋渡しする。	例）概念の橋渡しをする。

CEFR が設定している言語運用能力獲得の各レベルを到達目標にするということは、CEFR が言語活動として捉えている諸領域についての運用能力を到達目標に含むことを意味する。ここに示された諸活動領域において、口頭コミュニケーションが重要な位置を占めている。この表に示された例には偏りがあるが、実際には受容活動として「聞いて理解する」、産出活動として「話す」、やりとり活動として「口頭でやりとりする」、仲介活動として「通訳する」といった活動が口頭コミュニケーションにあたる。特に A1、A2 レベルといった入門～初級段階においては、人間関係を作って維持するためのパーソナルなやりとりや、生活に必要なモノやサービスを得たりする目的遂行的やりとりにおいて、口頭コミュニケーションの比重が大きい。そのような言語活動を遂行するに必要な能力の獲得を可能とする学習環境を提供することが、言語教育プログラムには求められる。これは別に CEFR を前提とするまでもなく、多くの外国語教育プログラムに求められていることである。本稿で報告するスペイン語教育プログラムも、このような言語学習のあり方に関する基本的な考え方に基づいていた。

基本理念と到達目標が設定されれば、学習内容と評価方法も自ずとそれらに沿った形で設計することになる。口頭コミュニケーション能力の獲得を促すと思われる学習教材を用意した。また本稿冒頭で触れた各セッションに配置された２名の母語話者授業補助者との口頭コミュニケーション活動が教室活動の中で重要な位置を占めた。さらに理念と到達目標に適合した評価を行うために、文法や語彙に関する小テストや音読課題など他の方法と併せ、これら授業補助者によるスペイン語口頭コミュニケーション能力についてのパフォーマンス評価を実施した。

３．対象授業科目の概要

本稿で報告するのは 2018 年度から 2021 年度までの４年間に実施した初年度対象のスペイン語クラス（各年度２クラス）において行った評価活動についてである。各年度、前期・後期の２期に分けて実施された。受講者は前期・後期を通して同じクラスで受講する規則になっていた。１学期は 15 週間、１回 90 分の授業が各週２回実施された。各年度の対象クラスの受講者数は表２のとおりである。後期に進めない学生がいるため、前期と後期では人数が異なる。なお 2020 年度は、Covid-19 のため年度を通して全ての授業がオンライン形式となった。各セッション二人の授業補助者もオンラインで授業に参加した。2021 年度は教室での対面授業となったが、間隔を開けて着席したりするなど、従来行っていた授業形態とは異なる状況が続いた。

表２．各年度の対象クラスの受講者数

	前期	後期
2018 年度	A クラス 24 名，B クラス 27 名	A クラス 24 名，B クラス 27 名
2019 年度	A クラス 32 名，B クラス 28 名	A クラス 32 名，B クラス 27 名
2020 年度	A クラス 31 名，B クラス 27 名	A クラス 30 名，B クラス 26 名
2021 年度	A クラス 33 名，B クラス 26 名	A クラス 32 名，B クラス 25 名

４．授業補助者が参加した評価活動の概要

スペイン語母語話者授業補助者が参加した評価活動は以下のようなものであった。

1） 授業内でのやりとり活動でのスペイン語運用能力の観察結果
毎回の授業で数分間の個別あるいは小グループでの授業補助者とのやりとり練習を行い、各学生のパフォーマンスについて直後に逐次評価する。形成的評価にあたる。

2） 学期途中および学期末のスペイン語でのやりとりテストの評価
それまでの授業内でのやりとり活動を踏まえた総括的なやりとりテストを、個別あるいは小グループで行い、各学生のパフォーマンスについて直後に逐次評価する。到達度評価にあたる。

3） 学期途中および学期末のスペイン語プレゼンテーション・ビデオの評価
課題に沿った内容のスペイン語によるプレゼンテーションを各自教室外で録画し提出させ、それを評価する。到達度評価にあたる。

５．評者間信頼性検定の概要と結果

このようにして得られた評価結果のうち複数の評者による評価結果の一部について評者間信頼性検定を行った。評価結果を順位データに変換し、スピアマンの順位相関係数（Spearman's rank correlation coefficient）を算出した。なお、本稿で報告する研究では、学習環境を構成する様々な要素を、経年的に厳格に制御することはしていない。

以下、いくつかの評価結果について評者間信頼性検定結果を提示する。なお、X, Y の値が共に同値のケースが多いことから、値に一定の加工を行ってケース数の視覚化を行なっている。

1）2018 年度最終口頭試験（2019 年 1 月実施）

最終口頭試験は、母語話者授業補助者と 1 対 1 でスペイン語でやりとりする形態で行った。一人の授業補助者とのやりとりは 2 ～ 3 分間で、それまでに授業内で行ってきたやりとり練習の内容をカバーする質問が授業補助者からなされ、これに応答していく形をとった。

評価にあたっては、時間的制約を考慮して詳細なルーブリックを用意することはせず、1)正確さ、2)流暢さ、3)創造性の 3 項目について 5 段階評価を求めた。なお 3 番目の「創造性」という項目は産出活動の評価で用いられる「複雑さ（complexity)」の要素を分かりやすく言い換えようとしたものである。

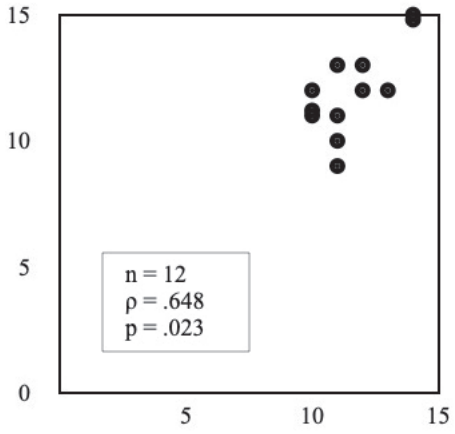

図 1 - 1

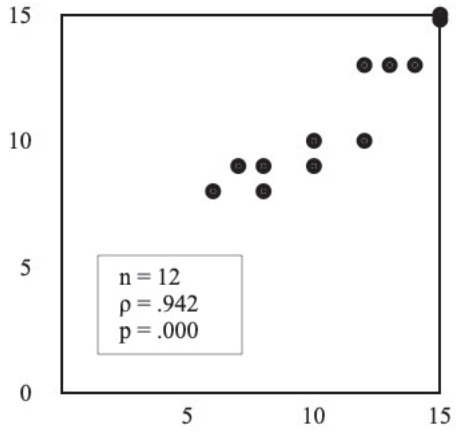

図 1 - 2

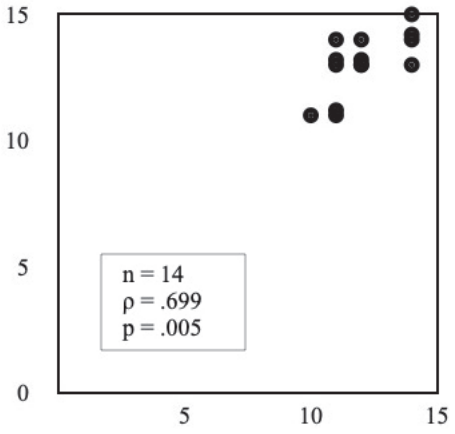

図 1 - 3

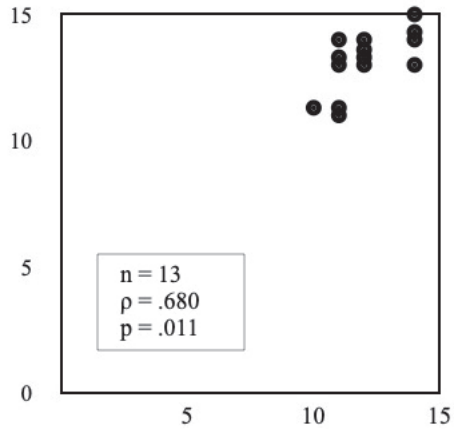

図 1 - 4

図１−１と図１−２は、2018 年度Ａクラスを２グループに分け、それぞれ２名の授業補助者が個別に面接し、その際のやりとりパフォーマンスをその場で採点した結果について示している。図１−３と図１−４は、2018 年度Ｂクラスについての同様の採点結果である。いずれの場合も有意な相関が見られ、図１−２に示したケースでは相関係数 ρ が 0.94 を超えており、かなり高い相関が見られた。他の３つのケースでは相関係数 ρ は 0.6 〜 0.7 の範囲で一定の相関が見らる結果となった。この課題では授業補助者の質問を理解して適切に応答することが求められ、授業補助者によってスペイン語発話の速度や発音の仕方に違いがあるなかで、概ね許容範囲の評者間信頼性が得られていたといえる。

２）2019 年度後期スペイン語プレゼンテーション・ビデオ（2020 年１月実施）

ここでは 2019 年度学年末に実施したスペイン語プレゼンテーション・ビデオの評価結果について示す。

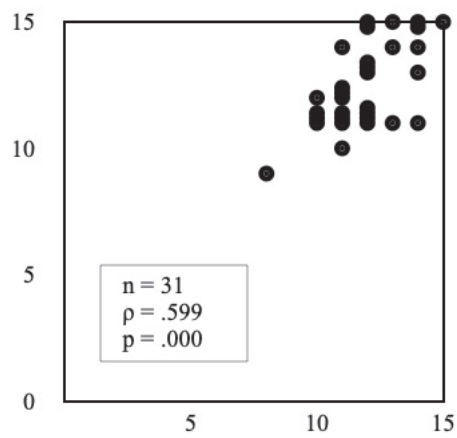

図２−１

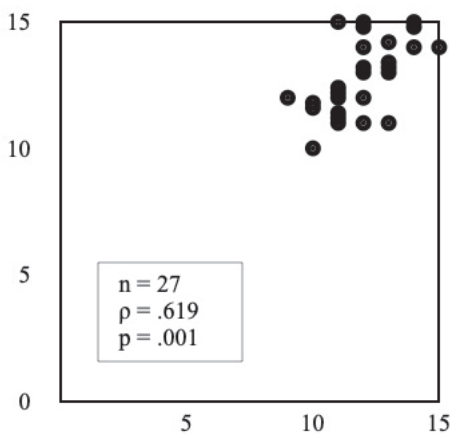

図２−２

図２−１と図２−２は、2019 年度Ａ・Ｂ２クラスの受講学生が教室外で一定の課題に沿ってスペイン語で語る様子をビデオに収録して提出したものをそれぞれ２名の授業補助者が採点した結果について示している。いずれの場合も相関は有意であり、相関係数 ρ は 0.6 前後であった。係数値は必ずしも高くないが、散布図から明らかなように満点の評価を得たケースが少なくなく、いわゆる天井効果が生じており、また著しい外れ値も見られないことから、概ね許容可能な評者間信頼性が得られていたと考えられる。

３）2020 年度後期学期末スペイン語プレゼンテーション・ビデオ（2021 年１月実施）

2020 年度の授業実施は、Covid-19 の感染爆発により極めて困難な状況に置かれた。結果的に年間を通じて全てのクラス・セッションをオンラインで実施した。自宅などからオンラインで参加する受講者を４名内外の小グループに分け､各グループのオンライン･セッションに授業補助者が一定時間ごとに順番に入る形でやりとり練習を行なうなど、口頭コミュニケーション能力の獲得を目指したが、さまざまな制約が生じた。

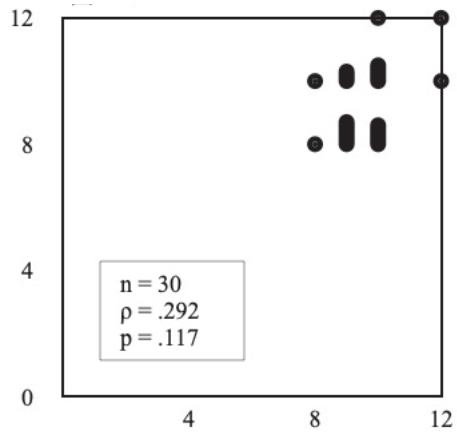

図３－１

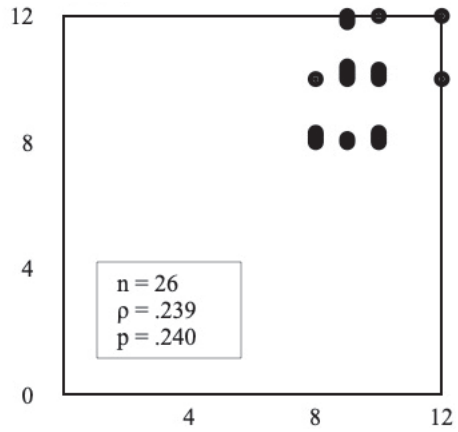

図３－２

図３−１と図３−２は、前年度と同様にＡ・Ｂ２クラスの受講学生が教室外で一定の課題に沿ってスペイン語で語る様子をビデオに収録して提出したものを、それぞれ２名の授業補助者が採点した結果について示したものである。授業補助者たちもいろいろなストレス下にあったこと、大きな制約下でやりとり練習時の評価を A+, A, B, C の４段階評価に簡略化していたことなどから、ビデオの評価も同様の４段階評価で行った。相関係数の算出は、４段階を A+=12 点、A=10 点、B= ８点、C= ６点に換算して行った。その結果、相関の有意性は却下された。ただし、散布図からはある程度の相関はあったように窺える。いずれにしてもパフォーマンス評価においては、最低限の条件を確保することが不可欠であることを痛感させられる結果となった。

４）2021 年度後期学期末スペイン語プレゼンテーション・ビデオ（2022 年１月実施）

2021 年度は、当該授業は教室での対面形式に戻った。ただし Covid-19 の影響は続いており、外国語の授業での口頭練習にあたっては、感染が生じないように十分注意するようにといった指示がなされたりした。紙に印刷した教材や小テストの使用が憚られ、2019

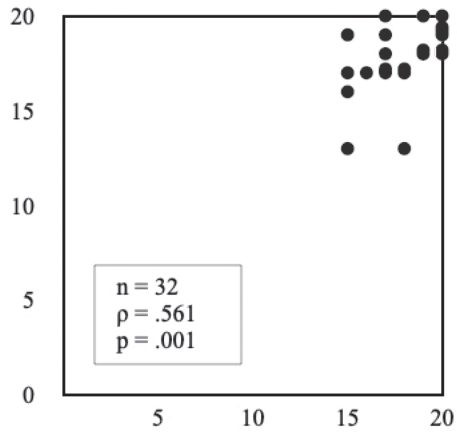

図４－１

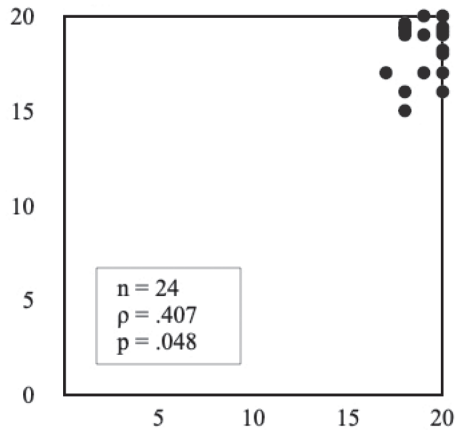

図４－２

年度まで実施していた小テストの比重が相対的に大きな授業形態に戻ることはできず、授業補助者とのグループでのやりとり活動とそれに向けたグループでの準備活動に比重を置いた授業形態が続いた。

前ページ図４−１と図４−２は、前年度までと同様にＡ・Ｂ２クラスの受講学生が教室外で一定の課題に沿ってスペイン語で語る様子をビデオに収録して提出したものを、それぞれ２名の授業補助者が採点した結果について示したものである。なお当該年度は以前用いていた 1)正確さ、2)流暢さ、3)創造性の３項目に加え、4)全体的な印象、という評価項目を設定した。各項目５段階評価、全体で最大 20 点という設定である。相関係数を算出した結果、いずれのクラスでも有意な相関が確認され、Ａクラスの場合での相関係数 ρ は 0.56、Ｂクラスの場合での相関係数 ρ は 0.41 となった。いずれも 2018 年度、2019 年度に比べて相関の度合いが弱くなっており、信頼性が相対的に低く算出される結果になってしまった。評価基準に関する意思疎通が十分でなかったこと、新たに加えた評価項目が具体性に欠け、却って天井効果を高めたり主観的判断の介入を強めることに繋がったかも知れないこと、などが原因として考えられる。

６．考察と今後の展望

口頭コミュニケーション能力についてのパフォーマンス評価は、テストの設計と実施に関して考慮すべき要素、すなわち妥当性（validity）・信頼性（reliability）・有益な波及効果（beneficial backwash）・実行可能性（feasibility）のうち、「妥当性」と「有益な波及効果」についてはかなりの程度確保することが可能である。本稿で報告した事例についても、エビデンスを提示することは難しいが、学習プログラムが掲げる口頭コミュニケーション能力の獲得という重要な目標に沿った評価が実施されたことで、受講学生の多くがそのような目標と評価形式に合わせた学習活動を行ったと思われる。

他方、「信頼性」と「実行可能性」の間には相反関係が生じやすい。本稿で報告してきたように、Covid-19 の感染拡大という外的要因の結果とはいえ、評価の実施にあたって最低限必要な態勢を整えなかった場合には、信頼性の確保が難しくなる。一方で、さほど厳密な評者間信頼性確保の手段を講じなくとも、一定の信頼性の実現は可能である。本稿で報告対象とした期間中も、評者間で事前に同じ評価対象について評価し、その結果をすり合わせるといったことを授業開始前の短時間に行ったりした。そのようなことの繰り返しを行なっていくことが重要であるが、現実にはさまざまな時間的制約があり、実現が容易ではないことも多い。

本稿で報告した事例では、パフォーマンス評価の信頼性係数値が必ずしも高くない場合もあった。筆者らの実施した授業では、評価には常に誤差が伴うものであることを考慮して、従来から評価の方法をできるだけ多様化し、一つの評価方法の結果の誤差が全体的な評点に大きく影響しないようにしていたが、そのような配慮も重要である。

他方、評者間信頼性を確保するに一定の工夫が必要であるということは、一人の評価者のみによるパフォーマンス評価をそのまま使うことの危うさを示している。教師が目標言語の母語話者である場合にも、別に評者を確保することが考慮されるべきである。

そのような場合も含め、本稿で報告した事例のように２名の母語話者授業補助者を各回の授業に配置することそのものが、実行可能ではないという状況がほとんどかも知れない。

筆者らの場合は、従前から学内特別予算を確保したり、今回のように科学研究費補助金の配分を得るなどして実行可能となった。評価活動も含めた外国語教育の教育経費について、学内学外各レベルでの理解を拡げ、資金面での実行可能性を確保していくことも、日本の大学における外国語教育に課せられた課題の一つといえよう。

付記：本研究は JSPS 科研費 18K00821 の助成を受けたものです。

参考文献

Council of Europe（2001）, *Common European Framework of Reference for Languages: Learning*, *Teaching*, *Assessment*, Council of Europe, Modern Language Division, Strasbourg, France and University of Cambridge Press, Cambridge, United Kingdom, available at https://rm.coe.int/1680459f97r 吉島茂・大橋理枝 訳編（2014）．外国語教育Ⅱ－外国語の学習、教授、評価のためのヨーロッパ共通参照枠－，追補版，朝日出版社

Council of Europe（2020）, *Common European Framework of Reference for Languages: Learning*, *Teaching*, *Assessment* – Companion Volume, Council of Europe Publishing, Strasbourg, France, available at www.coe.int/lang-cefr

Hughes, A.（2003）. *Testing for Language Teachers*. 2nd ed. Cambridge University Press. アーサー・ヒューズ著．静哲人訳.（2003）. 英語のテストはこう作る，研究社

小泉利恵，印南洋，深澤真編.（2017）. 実例でわかる英語テスト作成ガイド，大修館書店

根岸雅史 .（2017）. テストが導く英語教育改革，三省堂

望月昭彦，深澤真，印南洋，小泉利恵編著.（2015）. 英語 4 技能評価の理論と実践，大修館書店