

レポート評価におけるルーブリックの開発と その信頼性の検討

松 下 佳 代・小 野 和 宏・高 橋 雄 介
(京都大学) (新潟大学) (京都大学)

Development of a Rubric for Writing Assessment and Examination of Its Reliability

Kayo Matsushita, Kazuhiro Ono, Yusuke Takahashi
(Kyoto University) (Niigata University) (Kyoto University)

Performance assessments such as writing assessment have been generally thought to have low reliability, whereas they have high content and face validity. The purpose of this study is to develop a rubric for writing assessment and to examine the reliability of assessment with the rubric. In this study, we highlighted a writing assessment for the course “Study Skills at University” for the first-year students at the Faculty of Dentistry, Niigata University. This course was provided in the form of both lectures and seminars in the first semester. The outcomes of learning were assessed with student essays and presentations. We developed “Cross-Model Rubric,” which comprises six dimensions with three levels, to articulate logical thinking, problem solving and communication as the desired outcomes of this course. One of us has given a class of academic writing, using the rubric developed in this study as one of teaching materials, and three raters including the lecturer assessed student essays ($N=64$). With the goal of assessing the reliability, we used generalizability theory to estimate the contributions of student, rater, tasks, and their interactions to total variance. The generalizability coefficient, which is analogous to the reliability coefficient in classical test theory, scored 0.62, which was acceptable but did not reach a desired value. The generalizability theory analysis found that the writing assessment should be administered by more raters to achieve sufficient reliability by distinguishing individual students. In terms of feasibility of assessment, however, we propose that the rubric should have more levels of dimensions and the descriptors should be modified to appropriately assess student achievement.

〔キーワード：ルーブリック，信頼性，レポート評価，パフォーマンス評価，一般化可能性理論〕

1. 問題と目的

(1) 学習成果の評価への注目

学生は大学教育を通じて何を学んでいるのか。大学教育における学習成果の評価に対する要請は、1990年代頃から世界的に強まってきたが、わが国では、2008年12月

に出された中央教育審議会答申「学士課程教育の構築に向けて」において、学士課程修了段階の学習成果としての「学士力」やアドミッション・カリキュラム・ディプロマに関するポリシーなどの概念が提案されたことによって、広く認識されるようになった。2012年8月に出された答申「新たな未来を築くための大学教育の質的転換に向けて」ではさらに、今後の具体的な改革方策として、各大学における「アセスメント・ポリシー」の策定

や、「学修行動調査」「アセスメント・テスト」「ルーブリック」「学修ポートフォリオ」などの手法を用いた評価があげられている。(なお、2012年8月の中教審答申では、大学設置基準にあわせて「学修」という表記が使われているが、本稿では、必ずしも、設置基準のいう授業時間内・外の学びに限定しないことから、また、不要な煩雑さを避けるためにも、引用以外では「学習」という表記を用いることとする。)

こうした高等教育政策の動向に対応して、大学教育の現場でも、さまざまな形での学習評価が試みられるようになってきた。しかし、単に学習評価を多用するのでは、学生の学習を評価(とりわけ試験)に向けたものに矮小化したり、教員の評価負担を増大させたりするなど、負の効果ももたらしかねない。「評価の時代」において、我々は、それに振り回されるのではなく、それをうまくコントロールするすべを身につける必要があり、そのためには評価方法の批判的吟味が不可欠である。

(2) 学習評価のタイプ

現在行われている学習評価は、〈直接評価か、間接評価か〉、〈心理測定学的パラダイムによるのか、オルタナティブ・アセスメントのパラダイムによるのか〉によって、図1のように分類することができる(松下, 2012)。

直接評価とは、学生の知識や能力の表出(何がどのようにできるか)を通じて学習成果を直接的に評価することであり、間接評価とは、学生の学習行動(どのように学んだか)や学習についての自己認識の報告(何を学んだと思っているか)を通じて、学習成果を間接的に評価することである(Palomba & Banta, 1999; 山田, 2012)。一方、心理測定学的パラダイムとは文字通り、心理測定学に基礎をおく測定・評価の考え方であり、オルタナティブ・アセスメントとは、標準テストに代表される心理測定学的な評価方法を代替したり、補完したりするようにデザインされたさまざまな評価方法の総称である(Hart, 1994; Gipps, 1994; 松下, 2012)。前述

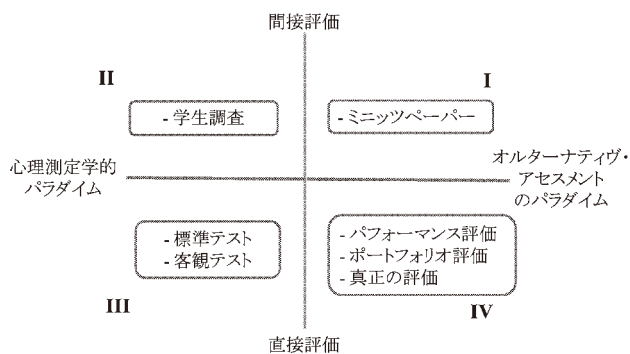


図1 学習評価のタイプ

(出典) 松下 (2012, p.81) を一部改変。

の手法のうち、「学修行動調査」はタイプII、「アセスメント・テスト」はタイプIII、「学修ポートフォリオ」はタイプIVにそれぞれ該当することがみてとれよう。

(3) ルーブリックを用いた評価

本稿で検討対象とするのは、タイプIVに含まれるパフォーマンス評価とその評価基準としてのルーブリックである。パフォーマンス評価とは、「ある特定の文脈のもとで、さまざまな知識や技能などを用いながら行われる、学習者自身の作品や実演(パフォーマンス)を直接に評価する方法」のことである(松下, 2012, p.76)。日本では、パフォーマンス評価の研究は初等・中等教育の方が先行しているが、学習者の作品や実演に基づく評価自体は、大学教育においても広く行われている。例えば、医療教育や教員養成などの分野での実技、芸術教育分野での演奏や美術作品、PBL (Project Based Learning) での製作物、さらには、レポート・卒業論文や口頭試問などを通じて行われる評価などである。だが、従来、往々にして、こうした評価は、個々の教員の主観にゆだねられたままであった(松下他, 2012)。

ルーブリックとは、パフォーマンス(作品や実演)の質を評価するために用いられる評価基準のことであり、1つ以上の基準(次元)とそれについての数値的な尺度、および、その内容を説明する記述語からなる(田中, 2003)。いいかえれば、パフォーマンスの質を多面的・段階的に評価するための評価基準である。近年、さまざまな組織や個人によって膨大なルーブリックが蓄積されつつあるが(濱名他, 2012)、なかでも注目されるのは、AAC&U (American Association of Colleges & Universities: アメリカ大学・カレッジ協会)のVALUE (Valid Assessment of Learning in Undergraduate Education: 学士課程教育における妥当な学習評価)プロジェクトである(Rhodes, 2010; 森, 2011; 飯吉, 2011; 吉田, 2011)。VALUEプロジェクトでは、標準テストに対抗・代替するものとして、eポートフォリオとルーブリックを組み合わせた評価方法が提案され、教養教育に関わる15の領域(探究と分析, 批判的思考, 創造的思考, 文章コミュニケーション, 口頭コミュニケーション, 読解, 量的リテラシー, 情報リテラシー, チームワーク, 問題解決, 市民参加, 異文化知識・能力, 倫理的推論, 生涯学習の基礎とスキル, 統合的学習)でルーブリックが共同開発された。それが「VALUEルーブリック」である(全訳は松下(2012)を参照)。例えば、「文章コミュニケーション」のルーブリックは、「文章作成の文脈と目的」「内容の展開」「ジャンルと学問分野の約束事」「資料と根拠」「構文と技法を操ること」の5つの基準

(次元)と「キャップストーン(4)」「マイルストーン(3, 2)」「ベンチマーク(1)」「ベンチマークに満たない(0)」の5つのレベルで構成されている。各数字は大まかには学年に対応しており、このルーブリックを用いた評価結果を通じて、学士課程教育による学生の成長、裏返せば、各大学の教育力(「付加価値(value-added)」)が測定できるようにデザインされている。また、VALUEルーブリックはいわばメタ・ルーブリックであり、それを各大学の文脈に応じてローカライズするという方法をとることによって、大学をこえた共通性と大学間の多様性の両立が図られている。

(4) 信頼性という問題

さて、一般に、測定・評価に求められる最も基本的な要件として、妥当性(validity)と信頼性(reliability)をあげることができる。妥当性は意図していた内容や構成概念を実際に測定しているか、一方、信頼性は正確に安定して測定しているか、を表す概念である。パフォーマンス評価では、妥当性、なかでも内容的妥当性や表面的妥当性は担保しやすいが、信頼性は担保しにくい。そこで、信頼性を担保するための一つの方法として、ルーブリックが用いられるのである。しかし、ルーブリックを用いるだけで、信頼性が担保される保証はない。にもかかわらず、VALUEプロジェクトの報告が特集されている*Peer Review*誌(2009年11巻1号, 2011年13巻4号)をみても、信頼性の検討がなされている例は少なく(Finley, 2011)、なされていても、その検討はなお不十分である。

信頼性という概念はもともと心理測定学の概念であり、オルターナティブ・アセスメントのパラダイムに立つ論者の中には、信頼性という概念の再考を主張する者もある。例えば、ギップス(Gipps, 1994)は、信頼性に代えて、採点者間の評価の一貫性を基礎とする「比較可能性(comparability)」という概念を提案している。ギップスの主張は、量的研究とは異なる質的研究の評価基準を構築しようとする議論(フリック, 2011)を基盤とするものであり、傾聴に値する。しかしながら、大学のように多様な教育観・評価観をもつ教員が集まっているところで、評価を公的なものとして用いていくためには、オルターナティブ・アセスメントのパラダイムに立つパフォーマンス評価においても、信頼性の要件を満たすことが望ましい。これは、理論的裏づけや信頼性の担保を伴わないルーブリックの流行が懸念されるなかで、学習の質を評価するためのよりすぐれたルーブリックを開発する上で必要な作業でもある。

信頼性係数の推定方法は、古典的テスト理論に基づく

方法が一般的であるが、より洗練されたものとして、一般化可能性理論に基づく方法がある(Brennan, 1997, 2001; 池田, 1994; Shavelson & Webb, 1991)。一般化可能性理論に基づく分析は、どの程度の評価観点数や評定者の人数であれば、どの程度の一般化可能性係数が得られるのかについてシミュレーションを行い、効果的に評価を行うことのできる計画を立てるための情報を与えてくれるものであり、評価改善に対する示唆を得やすい。

(5) 本研究の目的と構成

以上から、本研究では、パフォーマンス評価の一形態としてのレポート評価においてルーブリックを開発し、そのルーブリックを用いた評価の信頼性について一般化可能性理論に依拠しながら検討することを目的とする。その際、新潟大学歯学部初年次教育科目「大学学習法」におけるレポート評価を事例として用いる。以下では、まず、新潟大学でのレポート評価の実際、およびルーブリック開発の手続きについて述べる。次に、3名の教員によるレポート評価の結果の信頼性を、一般化可能性理論に基づく方法によって検討する。最後に、その結果をふまえて、ルーブリックを用いた評価を改善するための方法を提案する。

2. 方法

(1) 新潟大学初年次教育科目「大学学習法」とパフォーマンス評価

新潟大学歯学部では、知識基盤社会への移行に伴い、学士課程教育を歯科医療従事者としての生涯学習の最初の段階と位置づけ、問題解決能力の育成を重視し、その後続く大学院や実社会での学習のなかで、専門性を主体的に向上させうる人材を養成することとしている。この基本方針のもと、学生に自立・自律した学習者としての基礎を身につけさせるために、初年次教育として大学学習法を実施している。初年次の前期に新入生60名を対象に、15週30回で開講し、担当教員分担執筆による自作のテキストを使用して、図書館の利用方法、コンピュータを用いた情報検索や通信の基本技術、レポートの書き方やプレゼンテーションの仕方などについて講義・演習を行っている。また、講義・演習での知識や技能をもとに、レポートを作成し、その内容をプレゼンテーションするよう課題を課している。

この授業は、レポート作成およびプレゼンテーションという学習活動を通して、学生に問題解決能力、論理的思考力、表現力を身につけさせることを最終的なねらいとしているが、これまでは主に定型的な文書作成や口頭

発表のスキルを評価し、教育目標と評価の齟齬を認識しながらも、高次の統合的な能力をいかにして直接評価すべきか対応に苦慮していた(小野他, 2011)。どのような評価方法を用いるかは、教員が本当に重視しているものは何であるかを暗黙のうちに学生に伝えるといわれており(松下, 2007)、把握したい能力を直接的に反映する評価方法を用いなければ、誤ったメッセージを学生に与えることにもなりかねない。学生のモチベーションに影響し、ひいては発達を阻害するおそれさえある(Gipps, 1994)。

そこで2011年度から、大学学習法にパフォーマンス評価を導入すべく取組を開始した。パフォーマンス評価は個々の知識や技能の評価に重点を置く従来のテストとは異なり、現実的な状況で知識や技能を使いこなせる能力を評価することができる点に特徴がある(Hart, 1994)。このような能力は直接測定できないことから、要求に対するパフォーマンスを観察することにより推論することではしか把握できない。大学学習法にパフォーマンス評価を導入することによって、これまで評価が難しかった問題解決能力、論理的思考力、表現力をレポートにより可視化し、そこからループリックで能力を読み解こうと試みた(図2参照)。

(2) ループリック開発と十字モデル・ループリックの特徴

今回のループリックの開発にあたっては、まず著者の一人が素案を考え、大学学習法担当教員とともに2011年度の学生レポートを採点し、その結果について話し合いながら修正を重ねて作成した(表1参照)。このループリックは、「知識・理解」「問題発見」「情報検索」「論理的思考と問題解決」「文章表現」の5つの観点をもち、レポー

トの構成要素を網羅するものであったが、把握したい問題解決能力や論理的思考力とはどのような能力であるかという定義について十分な検討がないまま作業が進められたこともあり、この曖昧さから能力と観点の関連が不明確であった。また、「論理的思考と問題解決」はカテゴリーが大きすぎ、採点結果を学生指導に結びつけることが難しく、教育的利用という点でも問題がうかがえた(小野他, 印刷中)。

そこで、「十字モデル」(牧野, 2008)を理論的根拠として、問題解決能力や論理的思考力の定義を明確にし、ループリックの観点を整理することにした。「十字モデル」によれば、問題解決は、問題－主張－結論からなる過程(縦軸)としてとらえられ、また、論理的思考は、自分の意見を主張するために、その根拠と根拠を支持する具体的な事実やデータを示しつつ、同時に、対立する意見に対する論駁を行う過程(横軸)としてとらえられる。問題解決や論理的思考をこのように把握した上で、十字モデルのそれぞれの構成要素に一対一に対応する観点を設けることにより、理論的裏づけを備え、個々の学生レポートの強みや弱みを分析できる「十字モデル・ループリック」へと発展を図った(表2参照)。

(3) ライティング指導とレポート評価の実施

2012年度の大学学習法で、十字モデル・ループリックを用いて学生レポートを評価した。

新潟大学歯学部には歯学科と口腔生命福祉学科の2学科があり、2012年度の大学学習法受講者は歯学科40名、口腔生命福祉学科24名の計64名であった。

レポートの書き方を3回にわたり講義した最後に、レポート課題を示すとともに、十字モデルを用いて問題解決と論理的思考とは何かについて解説した。また、印刷

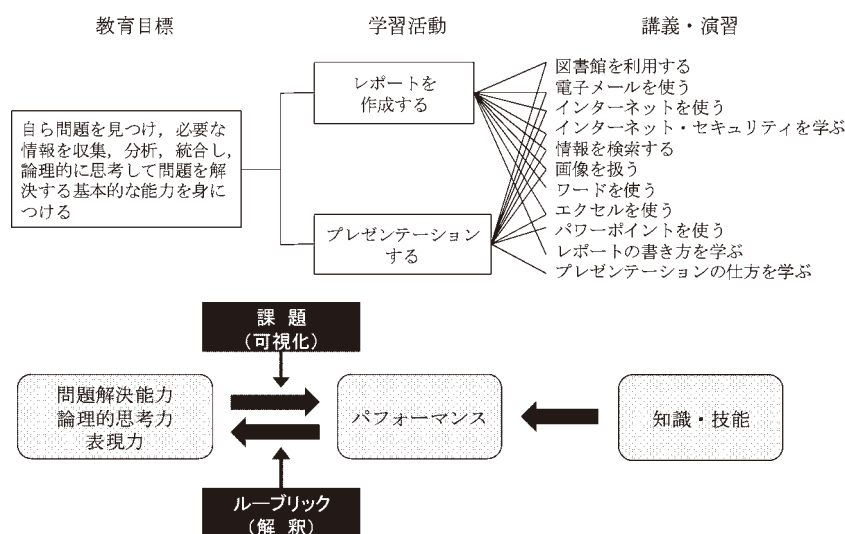


図2 大学学習法とパフォーマンス評価の関係

表1 レポート評価のルーブリック（2011年度）

観点	知識・理解	問題発見	情報検索	論理的思考と問題解決	文章表現
観点の説明	日本が直面する社会問題は、さまざまな要因が複雑に関連していることを理解する。	与えられたテーマに意欲的に接するなかで、疑問や問題を見いだす。	問題解決に必要な情報を収集し、その信頼性を吟味する。	収集した信頼できる情報をもとに、根拠ある意見を導きだす。	収集した情報と自分の意見を、結論にいたる過程を含めて、的確に表現する。
レベル3	自分が選択した話題について、多面的・多角的に検討し、その本質を理解している。	与えられたテーマから自分で話題を設定し、その問題を取り上げた理由や背景について述べている。	必要な情報を書籍およびウェブサイトなどから収集し、レポートの最後に出所を確認できる形で引用文献として記載している。	収集した情報を相互関連的にとらえ、分析的・実証的・論理的に考察し、根拠ある意見を導きだしている。	話題の選択から結論にいたる論理的な組み立て、記述の順序、パラグラフの接続が整っており、おおむね首尾一貫した文で書かれている。
レベル2	自分が選択した話題について、多面的・多角的に検討しているが、その本質の理解は不十分である。	与えられたテーマから自分で話題を設定しているが、その問題を取り上げた理由や背景の内容が不十分である。	必要な情報を書籍およびウェブサイトなどから収集しているが、その収集が不十分、あるいは信頼性の低い情報が散見される。	自分なりの意見を導きだしているが、収集した情報、あるいは考察が不十分で意見の根拠が薄弱である。	話題の選択から結論にいたるアウトラインはおおむね整っているが、記述の順序やパラグラフの接続に問題がある箇所が散見される。
レベル1	自分が選択した話題について、一面での理解にとどまっている。	与えられたテーマから疑問や問題を見いだせず、自分なりの話題が設定されていない。	情報の信頼性が批判的に吟味されていない。あるいは、レポートの最後に引用文献が記載されておらず評価できない。	情報収集のみで、自分の意見がない。あるいは、意見と感想の混同がみられる。	話題の選択から結論にいたる過程の記述が曖昧で、論理の流れを理解できない。
留意事項			信頼できる情報とは、大学、公的機関、学会、各種団体、新聞などの情報をさす。一方、信頼性の低い情報とは、作者や所属が書かれていないものや個人のブログなどの情報をさす。	根拠ある意見とは、信頼できる情報をもとに、論理的に考えて導きだされた意見であるが、情報源が示されていないと、信頼できる情報に基づいているか判断できない。したがって、引用箇所に肩番号がついていない場合は、評価は「レベル2」か「レベル1」となる。	文章構成を重視し、基本的に個々の文や表現の稚拙さなどは評価しないが、あまりに長い文や主語と述語の不一致など読者に誤解を与えるおそれのある文が多い場合は、評価は「レベル2」か「レベル1」となる。
			必ずしも書籍からの情報収集は必須ではない。	意見の有用性や実現可能性などの価値判断は、教員個人により異なるため評価に含めない。	

した十字モデル・ルーブリックを学生に配布し、レポートの評価基準について説明した。レポートのテーマは、現在の日本社会が直面している「エネルギー問題」と「大学の秋入学問題」の2つとし、学生数が同じになるように、学科や性別に関係なく、それぞれのテーマに学生を無作為に振り分けた。

学生は教員が提示した大きなテーマから、レポートで取り扱う具体的な話題を自分で設定し、授業時間外に主体的に調査・研究を行い、何かしらの意見や結論を導き、1か月間をかけてレポートを作成した。そして、その内容をプレゼンテーションの授業で発表し、そこでの質問や意見を参考にして、概要400～600字、本文2800～3200字のレポートを完成させた。なお、学生6名に1名のティーチング・アシスタントを配置し、必要に応じて学習支援を受けられるよう配慮した。

レポートの評価は、著者の一人も含め、大学学習法において特にライティング指導に関与した3名の教員で行った。評価に先立ち、十字モデル・ルーブリックに関して共通理解を得るために、ルーブリックの記述語とその意味するところを全員で確認した。また、2011年度の学生レポートのうち、特徴的な3本を採点事例として評

価者に配布した。これらのレポートは、著者らが十字モデル・ルーブリックで評価し、採点結果がよいもの（合計得点16）、中程度のもの（合計得点11）、悪いもの（合計得点9）の3本で、それぞれ、「表現ルール」は守られていないものの、他の観点はすべて「レベル3」のもの、「異なる意見の検討」を除くすべての観点が「レベル2」のもの、自分の「意見と結論」がなく、選択した話題について単に調査結果が述べられているものである。評価基準に関する話し合いの後、評価者3名は相互に相談することなく独立して、2週間かけて学生レポート64本を評価した。なお、2名の評価者は、十字モデル・ルーブリックによりレポートを評価するのは今回が初めてであった。

十字モデル・ルーブリックによって、学生はアカデミック・ライティングのポイントを理解できるようになり、一方、教員の側も学生の問題解決能力、論理的思考力、表現力をこれまでより根拠をもって評価できるようになった。しかし、評価の信頼性については検討の余地があり、現時点で、採点結果を総括的評価に用いることは時期尚早であると考えられた。そこで、2012年度は2011年度までと同様に、プレゼンテーションの実施とレポー

表2 十字モデル・ルーブリック（2012年度）

観点	問題解決		論理的思考		文章表現	
	背景と問題	意見と結論	根拠・情報	異なる意見の検討	全体構成	表現ルール
観点の説明	与えられたテーマから疑問や問題を見いだす。	自分の意見をもち、その意見を支える根拠・情報、あるいは対立する考えを考慮し、最終的な結論を導きだす。	意見を支える根拠を述べ、根拠の真実性を立証する情報を明らかにする。	意見と対立する考えやその考えを支える情報を明らかにし、反論に対して自分の意見の妥当性を示す見解を述べる、あるいは反論の問題点を指摘する。	話題の選択から結論に至る過程を的確に表現する。	指定されたルールを守り、文法に合った正しい日本語を使用してレポートを作成する。
レベル3	与えられたテーマから自分で話題を設定し、その問題を取り上げた理由や背景について述べている。	意見を支える根拠とその具体的な情報、あるいは対立する考えや論破の妥当性を考慮し、最終的な結論を導きだしている。	意見の根拠が述べられており、かつ根拠の真実性を立証する信頼できる情報が示されている。	意見と対立する考えやその考えを支える情報が述べられており、かつ反論に対して自分の意見の妥当性を示す見解や反論の問題点を述べている。	話題の選択から結論に至る論理的な組み立て、記述の順序、パラグラフの接続が整っている。	・文法に合った正しい日本語を使用している。 ・引用箇所に肩番号をつけ、レポートの最後に出所を確認できる形で引用文献を記載している。 ・字数制限が守られている。
レベル2	与えられたテーマから自分で話題を設定しているが、その問題を取り上げた理由や背景の内容が不十分である。	意見や結論は述べられているが、意見を支える根拠とその具体的な情報、対立する考えや論破の妥当性の考察が不十分である。	意見の根拠は述べられているが、根拠の真実性を立証する信頼できる情報が明らかにされていない。あるいは、根拠・情報の内容が不十分である。	意見と対立する考えやその考えを支える情報が述べられているが、反論に対して自分の意見の妥当性を示す見解や反論の問題点に言及していない。	話題の選択から結論に至るアウトラインはおおむね整っているが、記述の順序やパラグラフの接続に問題がある箇所が散見される。	・おおむね文法に合った正しい日本語を使用しているが、誤字や脱字を含め誤りが散見される。 ・引用箇所に肩番号がついていないが、レポートの最後に引用文献は記載されている。 ・指定された字数を越えているが、2割程度である。
レベル1	与えられたテーマから疑問や問題を見いだせず、自分なりの話題が設定されていない。	自分の意見や結論が述べられていない。	意見の根拠が述べられていない。あるいは、根拠・情報の内容がきわめて不十分である。	自分の意見と対立する意見についてふれていない。	話題の選択から結論に至る過程の記述が曖昧で、論理の流れを理解できない。	・あまりに長い文や主語と述語の不一致など読者に誤解を与えるおそれのある文が多い。 ・レポートの最後に引用文献が記載されていない。 ・字数制限が守られていない。
留意事項	教員はレポート課題として大きなテーマを学生に提示する。学生はレポートで取り扱う具体的な話題を設定する。	発見した問題に対する解決策を提案できることが望ましい。	根拠は、意見に「なぜならば」で結ぶことができる。	反対意見は、「たしかに～である」という表現で包むことができる。		3つの条件をすべて満たさない場合は、下位のレベルになる。
			情報は、根拠に「たとえば」で結ぶことができる。	反対意見の論破は、「たしかに～である」に続いて、接続詞「しかし」で結ぶことができる。		今回のレポートは、概要400～600字、本文2800～3200字としている。
			信頼できる情報とは、大学、公的機関、学会、各種団体、新聞などの情報をさす。一方、信頼性の低い情報とは、作成者や所属が書かれていないものや個人のブログなどの情報をさす。			
評価者はそれぞれを独立した観点として評価する。たとえば、レポートの結論が明確でなく、「意見と結論」が「レベル1」と判断された場合でも、何かしらの結論が導きだせる情報がレポートに記載されていれば、必ずしも「根拠・情報」は「レベル1」にならない。						

トの提出をもって一定の学修を示したとみなし、単位認定を行うこととした。

(4) 一般化可能性理論に基づく方法によるレポート評価の信頼性の検討

前述の通り、信頼性とは測定結果に一貫性・安定性があることだが、時間的に安定であること（再検査信頼性）、測定項目間で一貫していること（内的整合性）、評定者間で評価が一貫していること（評定者間一貫性）など、信頼性の内実は非常に多義的である。古典的テスト理論に基づく方法では、この多義性を考慮に入れることはせず、

ただ単に、信頼性を全分散における真値の分散の占める割合と定義したうえで信頼性係数の推定を行う。これに対して、一般化可能性理論による方法は、信頼性係数の推定を古典的テスト理論ではなく、分散分析の枠組みを利用して行う（Cronbach, Nageswari, & Gleser, 1963）。この方法は、評定者や項目など評価課題の諸側面ごとに誤差を分離して、より洗練された形で信頼性係数を推定することにより、古典的テスト理論に基づく考え方では解決できないパフォーマンス評価に伴う問題点を克服可能である（池田, 1994；Shavelson & Webb,

1991)。

一般化可能性理論は、一般化可能性研究 (Generalizability study: G 研究) と決定研究 (Decision study: D 研究) という 2 つの段階から構成される。G 研究では、評価において生じる測定誤差に着目し、その測定誤差の原因となる測定に伴う変動要因の成分とその分散を推定することによって、それぞれの変動要因やそれらの交互作用がパフォーマンス評価に与える影響を検討し、一般化可能性係数を算出する。さらに、D 研究では、G 研究で得られた各分散成分の推定値を用いて、評価観点数や評定者数をどの程度用意すると、どの程度の一般化可能性係数が得られるのかという点についてシミュレーションを行い、効果的な評価計画を立てるために必要な情報を得る。

そこで、本研究では、信頼性係数の推定のためのより洗練された手法としての一般化可能性理論を用いて、学生の手いたレポートに対するパフォーマンス評価の信頼性を検証する。具体的には、レポート評価における学生・評定者・評価観点という 3 つの変動要因の主効果とそれらの要因の交互作用の分散成分の推定値を求め (G 研究)、それらの推定値から一般化可能性係数を計算して、さまざまな評定者数および評価観点数ごとに得られる信頼性係数の値を求め、より適切な評価設計について検討を加える。レポート課題の評価がどの程度安定的かという点について検討を行うための今回のデータは、64 名の学生 (persons, p) の書いたレポートを 6 つの評価観点 (tasks, t) から 3 名の評定者 (raters, r) が評価する三相データで、分析デザインは、それぞれの評定者がすべての学生のすべての評価観点について評価を行い、すべての要因がクロスする $p \times t \times r$ デザインである。

3. 結果

(1) G 研究の結果

G 研究の結果得られたそれぞれの要因の分散成分の推定値および百分率、そして一般化可能性係数を表 3 に示した。一般化可能性係数は、学生 (p) に関わる交互作用のみを誤差として数値の推定を行った。

それぞれの分散成分の推定値は、すべての評価観点と評定者による評価を平均したときにおける数値を表したものである。学生の主効果 (p) は学生の得点のばらつきを示し、この分散成分の全体に占める割合が大きいことが学生のスキルそのものの評価に直結する。評価観点の主効果 (t) は評価観点間でばらつきがあるかどうか、評定者の主効果 (r) は評定者間における評価のばらつきがあるかどうかを示す。学生と評価観点の交互作用 ($p \times t$)

表3 各分散成分の推定値と一般化可能性係数

要因	分散成分の 推定値	(%)
学生 (p)	0.065	13.87%
評価観点 (t)	0.054	11.70%
評定者 (r)	0.042	9.09%
学生×評価観点 ($p \times t$)	0.022	4.65%
学生×評定者 ($p \times r$)	0.077	16.47%
評価観点×評定者 ($t \times r$)	0.010	2.07%
学生×評価観点×評定者 ($p \times t \times r$)	0.196	42.15%
一般化可能性係数	0.62	

は評価観点によって学生の順位が入れ替わる程度 (学生ごとの評価観点の得手不得手)、学生と評定者の交互作用 ($p \times r$) は評定者によって学生の順位が入れ替わる程度を示す。また、評価観点と評定者の交互作用 ($t \times r$) は評定者ごとの評価観点に対するとらえ方や重要視する度合いの違いを示し、学生・評価観点・評定者の交互作用 ($p \times t \times r$) は文字通りすべての要因の交互作用と測定誤差を含むものである。一般化可能性係数は 0.62 であった。一般化可能性係数を含む信頼性係数を評価する絶対的な基準は存在しないが、経験則として 0.80 を超える値を示す場合には十分と判断されることが多いので、今回の値は十分に高いものとはいえない。

全要因のうち、学生・評価観点・評定者の交互作用の割合が最も大きかった。この要因 (誤差を含む) が大きいことは、それぞれの主効果やお互いの交互作用で説明できない部分が大きいくことを意味し、信頼性係数の値が高くない原因の一つと考えられる。次に、今回のレポート評価は学生間の個人差を十分にとらえているとはいえない (全体の分散のうち 13.87% しか占めていない)。これが信頼性係数の低さの最大の原因といえよう。また、一次の交互作用については、評価観点ごとに学生の評価が異なったり、評定者ごとに観点を重要視する度合いが異なったりすることは少ないという結果が示された (それぞれ全体の分散の 4.65%, 2.07%)。

(2) D 研究の結果

先の G 研究で得られた各分散成分の推定値を用いて、D 研究を行い、評価観点数や評定者数をどのくらい用意するとどの程度の信頼性係数が得られるのか数値の変化をシミュレーションして、効率的でかつ実践に即したパフォーマンス評価を行うために必要な情報を得た。図 3 より、評価観点をこれ以上増やしても効率的に信頼性を高めることにはつながらず、信頼性の担保のためには評価観点数ではなく評定者数を増加させたほうが効率的であることがうかがえる。また、0.80 以上の信頼性を担保

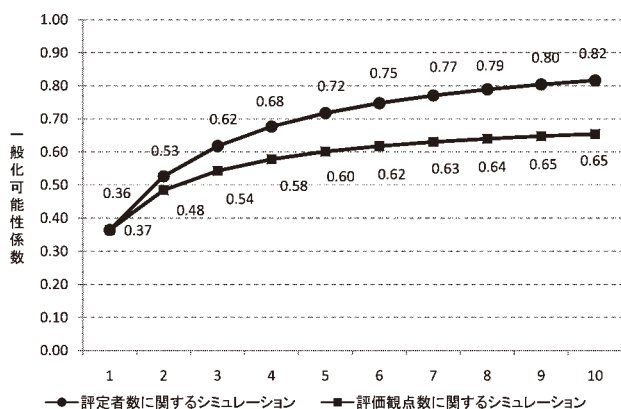


図3 一般化可能性係数のシミュレーション結果

するためには、評定者数の場合は今回の3名から9名へ、0.70以上の信頼性を担保する場合でも5名の評定者を用意する必要があることがわかる。

4. 考察

(1) ルーブリックを用いた評価の信頼性の検討

今回開発した十字モデル・ルーブリックを用いた評価の信頼性係数（一般化可能性係数）は0.62であり、一般に信頼性が十分と判断される値（0.80）には達していない。しかしながら、類似の領域での先行研究と比較した場合には、必ずしも低い値というわけではない。例えば、山西（2005）は、高校生の自由英作文評価において、一般化可能性係数が教員3名で約0.66であり、0.80を超える基準をめざすために必要な教員数は7名であったと報告している。同様に、宇佐美（2011）では、高校生の書いた小論文を4名の専門家が12個の評価観点から採点した場合、一般化可能性係数は約0.65であり、評定者を10名に増やしても0.80を超えなかった、とされている。さらに、大学生20名の口頭発表を3名が3つの観点から評価した場合の一般化可能性係数は、0.60を下回ったという研究結果もある（佐々木・村木，2005）。以上の先行研究結果をふまえると、今回の係数の値はあながち低いものとはいえず、先行研究とほぼ同等の結果を得られたとみることもできる。

(2) 評価の改善に向けて

とはいえ、評価の改善に向けてまだいくつかできることは残されている。図3に示したように、今回の評価結果についていえば、評価観点を増やすより評定者数を増やした方が、一般化可能性理論からすれば効率的ではある。しかしながら、評定者を3名から5名～9名へと増やすことは、評価の実行可能性（feasibility）を低下させ、現実的ではない。

今回の信頼性係数の低さの最大の原因は、学生の得点

の分散成分が全体に占める割合がわずか13.87%と低いことにあった。この得点の分散の小ささは、実際にレポートの質のばらつきが小さいというより、むしろレポートの質のばらつきが得点のばらつきに反映されていないことに起因すると推測される。この点に関して、一般化可能性理論においては分散成分の大きさを論じる必要があるため、ある程度のばらつきが期待できる方法でとられたデータを扱うことが望ましいことが、先行研究でも指摘されている（山森，2003）。

得点の分散をレポートの質のばらつきにあわせてより大きくする方法として考えられるのは、レベルの数を増やすとともに、各レベルの記述語をより学生の水準にそったものに修正することである。現在のルーブリックでは、3つのレベルが設定されており、加えて、レベル3の水準がかなり高い一方で、レベル1は、（2012年度前期の大学学習法のような丁寧なライティング指導を経た学生にとっては）低い水準になっているために、レベル2の幅が相当広がっている。レベル1の水準を、ライティング指導をふまえてもう少し引き上げ、VALUEルーブリックのように、レベル1に満たないものはレベル0とする、といった形に変えれば、より正確にレポートの質が得点に反映されるようになり、評定者もルーブリックにそった評価がしやすくなって、一貫性・安定性が高まると期待される。

今回のルーブリックは、「問題解決能力」「論理的思考力」「表現力」といったあいまいな概念を、学生の学習や教員の指導にとって理解可能なレベルにまで具体化することには成功したといえる。今後は、本研究の検討結果をふまえて信頼性を向上させることにより、教授・学習にとってより効果的なツールになるよう改良を重ねていきたい。

参考文献

- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Brennan, R. L. (1997). *Elements of generalizability theory* (Revised ed.). Iowa City: ACT Publications.
- Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, 16, 137-163.
- Finley, A. P. (2011). How reliable are the VALUE Rubrics? *Peer Review*, 13(4), 31-33.
- フリック, U. (2011)『新版 質的研究入門—人間の科学

- のための方法論—』(小田博志監訳) 春秋社。
- Gipps, C. V. (1994). *Beyond testing: Toward a theory of educational assessment*. London: Falmer Press. ギップス, C. V. (2001)『新しい評価を求めて—テスト教育の終焉—』(鈴木秀幸訳) 論創社。
- 濱名篤・土持ゲーリー法一・井上敏憲・松田岳士・山本秀樹・藤木清 (2012)「体系的なカリキュラム構築と学習成果の可視化のためのルーブリックの構築・活用」,『大学教育学会誌』34-2, pp.82-85.
- Hart, D. (1994). *Authentic assessment: A handbook for educators*. Menlo Park, CA: Addison-Wesley.
- ハート, D. (2012)『パフォーマンス評価入門—「真正の評価」論からの提案—』(田中耕治監訳) ミネルヴァ書房。
- 飯吉弘子 (2011)「学生のラーニングアウトカム向上のための教育実践と評価—多人数課題型学習効果の検証—」,『名古屋高等教育研究』11, pp.273-292.
- 池田央 (1994)『現代テスト理論』朝倉書店。
- 牧野由香里 (2008)『「議論」のデザイナー—メッセージとメディアをつなぐカリキュラム—』ひつじ書房。
- 松下佳代 (2007)『パフォーマンス評価』日本標準。
- 松下佳代・坂本尚志・小野和宏・高橋雄介・平山朋子・関田一彦 (2012)「学習成果の直接評価に向けて—パフォーマンス評価の可能性—」,『大学教育学会誌』34-2, pp.86-89.
- 松下佳代 (2012)「パフォーマンス評価による学習の質の評価—学習評価の構図の分析にもとづいて—」,『京都大学高等教育研究』18, pp.75-114.
- 森利枝 (2011)「高等教育機関の特性を活かした客観的指標の模索」,『文部科学教育通信』282, 16-17.
- 小野和宏・八木 稔・ステガロユ ロクサーナ・大島勇人・西山秀昌・八巻正樹・前田健康 (2011)「初年次教育としての大学学習法の効果と課題」,『日本歯科医学教育学会雑誌』27, pp.66-77.
- 小野和宏・西山秀昌・八木 稔・ステガロユ ロクサーナ・重谷佳見・山村健介・井上 誠・前田健康 (印刷中)「大学学習法へのパフォーマンス評価導入における実践的課題」,『新潟大学高等教育研究』1.
- Palomba, C., & Banta T. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco: Jossey-Bass.
- Rhodes, T. (Ed.) (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities.
- 佐々木典彰・村木英治 (2005)「口頭発表の評価における信頼性—一般化可能性理論を用いて—」,『教育情報学研究』3, pp.1-4.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.
- 田中耕治 (2003)『教育評価の未来を拓く—目標に準拠した評価の現状・課題・展望—』ミネルヴァ書房。
- 宇佐美慧 (2011)「小論文評価データの統計解析」,『行動計量学』38-1, pp.33-50.
- 山田礼子 (2012)『学士課程教育の質保証へむけて—学生調査と初年次教育からみえてきたもの—』東信堂。
- 山森光陽 (2003)「中学校英語科の観点別学習状況の評価における関心・意欲・態度の評価の検討—多変量一般化可能性理論を用いて—」,『教育心理学研究』51, pp.195-204.
- 山西博之 (2005)「一般化可能性理論を用いた高校生の自由英作文評価の検討」,『JALT Journal』27-2, pp.169-185.
- 吉田武大 (2011)「アメリカにおけるバリュールーブリックの動向」,『関西国際大学教育総合研究所研究叢書』4, pp.1-12.

謝辞

「十字モデル」について、直接、また懇切丁寧なご指導をいただきました、関西大学・牧野由香里氏に心より感謝の意を表します。

【付記】

本研究の一部は、平成23年度新潟大学授業改善プロジェクト「大学学習法へのパフォーマンス評価の導入——問題解決能力と論理的思考力を評価する——」,平成24年度新潟大学組織的教育プロジェクト「高次な能力を読み解く新しい評価方法の開発」,ならびに科学研究費基盤研究(C)「深い学習を促すパフォーマンス評価の開発——OSCE-Rを中心に——」(平成24~26年度,研究代表者:松下佳代,課題番号24530953)の支援を受けた。