研究論文

パフォーマンス評価における学生の自己評価・相互評価は妥当な評価に近づきうるか
―市民的オンライン推論能力を素材として―

長沼　祥太郎[1]・杉山　芳生[2]・澁川　幸加[2]・浅川　裕子[3]・松下　佳代[4]

（[1]九州大学教育改革推進本部・[2]京都大学大学院教育学研究科・[3]独立行政法人国際協力機構・[4]京都大学高等教育研究開発推進センター）

本研究の目的は、学習者自身によるパフォーマンス評価の評価結果は妥当な評価に近づきうるかどうかを明らかにすることである。本研究ではまず、「市民的オンライン推論能力」を素材として、パフォーマンス課題とルーブリックを開発した。先行研究を参考に、回答を客観的に判断可能な箇所を課題に組み込み、ルーブリックとの対応関係を明確化して採点しやすさを追究した。次に、ある私立大学の学生 90 名がパフォーマンス課題に取り組み、ルーブリックを用いて自己評価・相互評価を行った。分析では、これらの学生の自己・相互評価の結果を課題作成者による評価結果と比較し、それらの一致度（一致率・相関係数）を算出した。その結果、いずれも高い一致度を示した。本研究は、学習者自身の自己評価や相互評価は、評価方法や対象とする能力によっては、より専門的な鑑識眼を持った採点者と大きな齟齬なく採点できることを示した。このことにより、パフォーマンス評価の実行可能性を高める上で、学習者自身の採点結果を使用できる可能性が示唆された。

キーワード：パフォーマンス評価、実行可能性、市民的オンライン推論能力、自己評価、相互評価

1. 問題と目的

1.1. パフォーマンス評価への注目

2008 年の学士課程答申（中央教育審議会, 2008）および 2012 年の質的転換答申（中央教育審議会, 2012）を契機として、大学教育において授業で「教員が何を教えたか」だけでなく「学生が何を学んだのか」に大きく注目が集まるようになってきた。この流れの中で、「学生が何を知り、何ができるようになったか」を捉えるための教育評価の重要性が増している。そして、「実際に何ができるのか」という知識・技能の活用能力を見るための評価方法として、「パフォーマンス評価」が注目されている。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」（松下, 2012, p. 76）のことである。大学教育では、これまでにもレポートや実技を課すといった評価が行われてきたが、主観的な評価に陥りがちであった（松下, 2012）。一方で、近年のパフォーマンス評価論においては、より客観的な評価へと近づけることを意図して、後述の「ルーブリック」と呼ばれる評価基準表の使用が前提となっている。そのため、本稿では、「パフォーマンスをルーブリックにより解釈しようとする評価方法」を「パフォーマンス評価」と呼ぶこととする。

1.2. パフォーマンス評価の妥当性と信頼性

パフォーマンス評価では、学習者が「何ができるのか」を直接表出させるような「パフォーマンス課題」が用いられる。そのため、多肢選択式問題に比べて、特に内容的妥当性や表面的妥当性を担保しやすいと言われている（cf. Hart, 1994 田中訳, 2012）。

そして、パフォーマンスの評価を行う際に、信頼性を高めることが期待されているのがルーブリックである（西岡他, 2015）。これまでの大学教育では学生の作品や実演の評価は主に教員個々人の主観にゆだねられていた（松下, 2012）。そのため、多肢選択式問題を用いた評価では問題とならない「評価者間信頼性」、すなわち、別の人が採点しても同じ採点結果になるのかどうかがパフォーマンス評価では懸念されやすい。この懸念への対処が期待されるのが、「ルーブリック」である。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」である（松下, 2012, p. 82）。ルーブリックを用いることで、複数の評価者で採点した場合にも、ある程度の評価者間信頼性を得ることに成功している（例えば、

佐々木・村木, 2005；山西, 2005）。このように、ルーブリックの使用により、複数の評価者でもある程度の共通認識を持って学生のパフォーマンスを評価できるようになるのである（Reddy & Andrade, 2010）。

こうした研究により、信頼性と妥当性の基準を満たしたパフォーマンス評価を実現できることが示されてきた。しかしながら、文部科学省（2017）の調査によると、学部段階において課程を通じた学生の学修成果の把握を行っている354大学のうち、学修評価の観点・基準を定めたルーブリックを用いている大学は51校（14.4%）と、その使用率はかなり低い。これは教育課程レベルでの話であり、各授業レベルでの使用を示すものではないが、こうした数値からも、教育現場においてパフォーマンス評価の普及の現状はいまだ厳しいことが推察できる。

1.3. パフォーマンス評価の実行可能性

パフォーマンス評価が広がらない大きな要因として、松下他（2015）は、「実行可能性」の低さ（評価負担の大きさ）を指摘している。実行可能性とは、「利用可能な資源と時間の限度内で、評価対象としなくてはならない人数の子どもたち［学習者］を評価できるか」（田中他, 2011, p. 215, カッコ内は引用者）とされる。実行可能性について西岡他（2015）は、「教育現場における教育評価は、どれだけ信頼性と妥当性を備えていようとも、使える時間やリソースの範囲内で行えるもの（つまり実行可能性を持つもの）でなければ、それは絵に描いた餅にすぎない」と述べている。すなわち、パフォーマンス評価を広めるには、妥当性や信頼性のみならず、実行可能性の高さを満たした評価方法の開発が求められていると言えよう。

こうした要請に対して、大きく2つの方向性が提案されている。1つは、Darling-Hammond & Adamson（2010）が提案するように、課題のデザインや採点方法を効率的にすることである。この方法は、おそらく最も教育現場で受け入れられやすい汎用的な方法であろう。一方で、この方法の欠点としては、どれだけ採点が簡便になろうとも、教員が多忙な業務の中で採点の時間を確保することは容易ではなく、大人数クラスでの実施となればなおさら難しいということが挙げられる。

2つ目は、松下（2016）やBouzidi & Jaillet（2009）らが提案するように、TAや学習者自身といった、教員以外の評価者が採点を行う方向性である。特に学習者自身による採点として、自己評価と相互評価の利用が提案されてきた。これらは、教員の評価労力を必要としないため、教育現場における高い実行可能性を担保できると言える。ただし、これは教員による評価ではないため、採点結果自体を妥当なものとして見て良いのかという大きな問題が浮上する。採点結果が教員による評価結果とどの程度一致するかを慎重に検討する必要がある。学習者による評価と教員による評価結果の一致度が極端に低い場合は、学習者による評価結果をそのまま教育現場で教授改善や成績評価に用いることは到底できないため、真に実行可能性があると言うことはできない。すなわち、実行可能性の前提として評価結果の妥当性が必要なのである。

1.4. 自己評価と相互評価

自己評価や相互評価と教員による評価結果の一致度は、Ashenafi et al.（2017）のレビューに見られるように、これまで国外において多く研究されてきた（e.g., Campbell et al., 2001; Lin et al., 2001; Patri, 2002; Rudy et al., 2001）。表1は先行研究において報告された専門家と学生の評価の相関係数をまとめたものである。

表1　先行研究の専門家と学生の評価の相関係数

	自己評価	相互評価
Campbell et al.（2001）	.20～ .45	.03～ .73
Lin et al.（2001）	報告なし	.23～ .40
Rudy et al.（2001）	.19	.50
Patri（2002）	.46～ .50	.49～ .85
斎藤他（2016）	.076	報告なし

この表から明らかなように、相関係数には大きなばらつきが見られる。すなわち、自己評価や相互評価の結果を妥当な評価結果と一律に見て良いわけではないことが実証的に示されてきた。国内においても、同様の問題意識のもとで、いくつかの研究知見が示されている。例えば斎藤他（2016）は、歯学部の学生を対象にアカデミック・ライティング能力を評価するためのパフォーマンス課題を与え、学生の自己評価結果と教員による評価結果の相関を検討している。その結果、教員による評価結果と学生の自己評価結果の間に線形的な相関は見られず（r = .076）、加えて学生の自己評価のほうが教員による評価よりも得点が高くなる傾向があったと報告されている。また、木下他（2016）によるレポート課題を用いた研究では、ルーブリックを用いて11項目でレポートの評価を行った結果、どの項目においても教員と相互評価の結果の一致率は50%前後であったと報告されている。その理由として、学生がルーブリックを理解できていない、あるいは判断基準が曖昧であったことを挙げている。

このように、学生の自己評価や相互評価の質に関しては、国内では否定的な証拠の方が多い。実際、これまで自己

評価や相互評価は、主に学習内容の理解度、メタ認知能力といった認知領域や学習意欲、自律的動機付けといった情意領域に対する肯定的な影響の方が、国内においては注目されてきた（布施・岡部, 2010；河野他, 2014；生田目, 2004；西方, 2017；西岡他, 2015；尾澤他, 2004）。その一方で、学生による自己評価や相互評価の結果自体が妥当な点数と見なされることはほとんどなかった。

しかしながら、これらの学生の手による評価結果を妥当でないと判断するのは早計である。なぜなら、1.3. で述べた２つの方向性を組み合わせることで、学生の自己評価や相互評価の結果を、教員による評価結果という、より妥当な評価へと近づけることも期待できるためである。すなわち、解決へのアプローチとして、学習者自身でも評価しやすいように評価課題や採点方法を開発するという方向性がありうる。一方で、このようなアプローチを取り入れた試みは、管見の限り見当たらない。

1.5. 目的

そこで本研究では、採点しやすさを追究したパフォーマンス課題とルーブリックを用いた場合、パフォーマンス評価における学習者の評価結果は妥当なものとなりうるかを明らかにすることを目的とする。そのために、SHEG（2016）を参考に、「客観的に判断可能な箇所」を課題に組み込み、「ルーブリックとの対応関係」を明確化することで採点しやすさを追究したパフォーマンス評価を開発した。その後、パフォーマンス課題に学生に取り組ませ、ルーブリックを用いて採点させ、学生による評価結果（自己評価・相互評価）が課題作成者による評価結果とどの程度一致するかを実証的に検討した。なお、本研究では評価対象とする能力として、パフォーマンス評価と親和性が高いと考えられる以下の「市民的オンライン推論能力」を取り上げた。

2. 評価方法の開発

2.1.「市民的オンライン推論能力」とパフォーマンス評価

2016 年のアメリカ大統領選挙を主なきっかけとして「フェイクニュース」という言葉が市民権を得た。これを一因として、現在、オンライン空間に溢れかえる情報の質を判断する能力の育成が重視されている。この状況の中で注目が集まっているのが、「市民的オンライン推論能力（Civic Online Reasoning）」（McGrew et al., 2017; SHEG, 2016）である。市民的オンライン推論能力とは「社会的・政治的問題について、デジタルコンテンツを評価し、根拠づけられた結論を導く能力」と定義される（McGrew et al., 2017, p. 5）。この定義は、類似の構成概念である「情報リテラシー」の定義（例えば、国立大学図書館協会, 2015；山内, 2003）と重なるところもあるが、「社会的・政治的問題を対象としたオンライン上のコンテンツ」に対象を限定している点に特徴がある。これは、現代の情報化社会においてすべての一般市民が日常的に生活を送っていく上で必要な能力である。教養ある市民を社会に送り出すことを使命に持つ大学教育機関には、市民的オンライン推論能力の育成が求められているといえよう。そして、この能力の育成において、教育評価が重要であることは論を待たない。

市民的オンライン推論能力の評価は、パフォーマンス評価と非常に親和性が高い。なぜなら、多肢選択式問題によって、記事の質を判断する上での知識を有しているかどうかを尋ねる（例えば、オンライン記事の URL を見て公的な機関による信頼できる情報かどうかを同定する方法を知っているか問う）よりも、パフォーマンス評価によって、実際にオンライン上にある記事の情報が信頼できるかを判断させる方が、より能力を直接的に表出させることができると考えられるからである。

2.2. 市民的オンライン推論能力の下位概念

McGrew et al.（2017）では、市民的オンライン推論能力の下位に「a. 情報発信者の同定」、「b. エビデンスの質の評価」、「c. 他の情報の検索」という３つの能力を設定している。「a. 情報発信者の同定」では、記事はどのような個人あるいは団体により書かれたものであるかを同定する能力である。「b. エビデンスの質の評価」は、記事で書かれている主張を支えるエビデンス及びその論拠の質を評価できる能力を指す。

「c. 他の情報の検索」は、McGrew らが非常に重視している能力である。これは、提示された記事の信憑性を確認するために、その記事で書かれた内容に関連する他の記事を探すために外に読みを広げることができる能力である（McGrew らは「ラテラルな読み（lateral reading）」と呼んでいる）。McGrew らがこの３つ目の能力に重きを置くのは、ラテラルな読みをすることが情報の真偽を見極める専門職である「ファクト・チェッカー」が頻繁に行う行動であるためである。ファクト・チェッカーは、ある記事が自分のよく知らない内容であった場合、その記事を注意深く読み続けることはしない。彼らは一旦その記事を離れ、他の関連記事を探すことを通じて、元の記事の内容について学ぼうとするのである。この行動の理由は、「一見逆説的に見えるかも知れないが、これによりファクト・チェッカーはインターネットの強みを活かして、問題の記事を、より広いウェブという網の中で位置づけることができるのである」（McGrew et al., 2017, p. 8）と説明されている。さらに彼らは、その記事だけを注意深く読む（close reading）ことに比べて、ラテ

ラルな読みを少しでも行うことが情報の真偽を確かめる上で非常に大きな効果を生むとも主張している。また、このようなラテラルな読みは、「a. 情報発信者の同定」「b. エビデンスの質の評価」と独立ではなく、まさに情報発信者を同定したり、記事内のエビデンスの質の評価を行うために行われるべき行動であると言える。

よって、本研究においては「c. 他の情報の検索」を独立した下位概念ではなく、「a. 情報発信者の同定」および「b. エビデンスの質の評価」の高いレベルで見られる行動とみなして３つの下位概念を再整理し、関連づけることにした。

2.3. パフォーマンス課題とルーブリックの開発

本節では、オンライン推論能力の評価方法の開発過程を説明する。まず、２つの架空のオンライン記事（A、B）を作成した。記事 A では「高度外国人材にとって日本は魅力的である」、記事 B では「遺伝子組換え作物は健康によい」という主張がそれぞれ展開されている。これらは、人材の移動の流動性の高まり、そして人口増加による食糧難を迎えている現代の社会において、市民も考えなければならない社会的・政治的な問題である。記事内では「一つの主張」「主張の根拠」「記事執筆者情報」を示した。このオンライン記事を読み、書かれている主張の真偽をオンライン環境下において判断させるというのが今回のパフォーマンス課題である。なお、記事 A、記事 B の記事執筆者情報には、氏名と所属が示されている。両記事ともに氏名は第二・第三著者と同じ研究科に所属する大学院生２名の氏名を本人からの許可を得て使用し、所属は架空の機関の名前を設定した。これにより、「a. 情報発信者の同定」をするために記事執筆者を検索した際に、氏名と所属が一致する検索結果が得られないため、その記事の信憑性を疑問視できることを意図した。

次に、このパフォーマンスを可視化し、解釈するツールとして、次に説明する回答用紙およびルーブリックの作成を行った。これらの作成においては、SHEG（2016, pp. 8–14）を参考とし、「客観的に判断可能な箇所」を課題に組み込み、「ルーブリックとの対応関係」を明確化することで採点しやすさを追究した。SHEG（2016）の回答用紙では、オンライン上のある記事が広告かどうかを判断するという課題において、まずそれが広告か否かを正誤問題で問い、その上で自らが下した判断の理由を記述させる。すなわち、正誤問題という客観的に判断可能な箇所を組み込んでの採点が志向されているのである。この形式により、表２の記述語から読み取れるように、誤った選択肢をとっていれば、その時点で一番低いレベル「これから」を割り当てることができる。そして、正誤問題に正解していた場合に限り、自由記述が評価の対象となり、その質によりレベルが最終的に割り当てられる。この場合でも、評価する際に確認すべきレベルは２つのみ（「修得している」と「発展中」）であり、容易に評価できる。

本研究で開発した回答用紙の一部を図１に示した。図１の問①では、回答者は、自分がどのようにその記事の真偽を判断したのか、そのプロセス（と結論[1]）を回答用紙に書くように求められる。回答用紙の左側には、オンラインで提示したものと同じ記事が貼ってあり、まず主張の真偽を判断する上で着目した箇所に印をつけさせる。次に、その理由を印の近くに記述するよう指示した。問②では、「他の情報ソース」から情報を集めようとしていたかどうかを評価するために、「②記事にない情報を集めるために、他の web サイトの情報を見ましたか?」という問いに対して「はい」「いいえ」のどちらかで回答させた。問②で「はい」の場合には、問③で検索した記事の内容について書かせた。

この回答用紙に対応する形で、表３のルーブリックを作成した。これは、２つの記事 A、B の双方に対応するものである。このルーブリックでは、1）記事を読む際にある箇所に注目したかどうかが問①での印の有無によって即座に判断でき、2）問②で「いいえ」を選択した場合にはレベル 1、2、3 のうちのどれかが割り当てられ、「はい」を選択した場合に限り、レベル 4 の基準を満たすか否かが判断される。これらは、SHEG のルーブリックにおいて、客観的に判断可能なように二値的な正誤問題を取り入れていることを援用したものである。

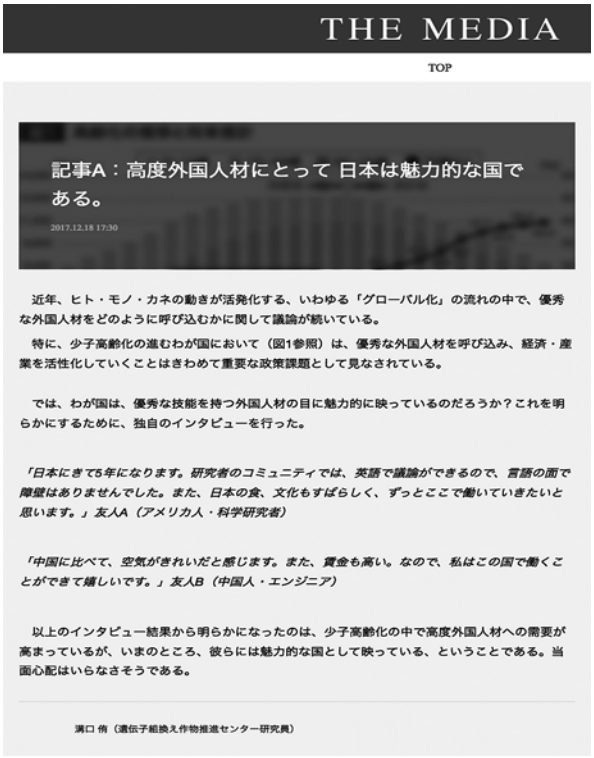
こうして作成したパフォーマンス課題（回答用紙含む）とルーブリックの内容及びデザインを、筆者らを含む教育評価を学ぶ大学院生、学部学生および教育評価の専門家計 10 名以上で繰り返し洗練し、表面的・内容的妥当性の向上を図った。

表２　SHEG（2016）の市民的オンライン推論能力を評価するルーブリック*（一部のみ抜粋）

レベル	修得している（mastery）	発展中（emerging）	これから（beginning）
記述語	学習者はこの項目が広告であるか広告でないかを正しく特定し、一貫した推論能力を示している	学習者はこの項目が広告であるか広告でないかを正しく特定しているが、推論能力が限定的、あるいは、一貫していない	学習者はこの項目が広告であるか広告でないかを正しく特定していない

* SHEG（2016, p. 18）より筆者らが一部修正の上訳出

① 記入例を参考にして、記事の主張の真偽に関し、**あなたが注目した箇所に印をつけ**（線を引く、〇で囲む、等）、その箇所をどのようにあなたの判断に利用したのか、具体的な理由があればそれを**印の近くに、記述してください。**



THE MEDIA

TOP

記事A：高度外国人材にとって 日本は魅力的な国である。

2017.12.18 17:30

近年、ヒト・モノ・カネの動きが活発化する、いわゆる「グローバル化」の流れの中で、優秀な外国人材をどのように呼び込むかに関して議論が続いている。

特に、少子高齢化の進むわが国において（図1参照）は、優秀な外国人材を呼び込み、経済・産業を活性化していくことはきわめて重要な政策課題として見なされている。

では、わが国は、優秀な技能を持つ外国人材の目に魅力的に映っているのだろうか？これを明らかにするために、独自のインタビューを行った。

「日本にきて5年になります。研究者のコミュニティでは、英語で議論ができるので、言語の面で障壁はありませんでした。また、日本の食、文化もすばらしく、ずっとここで働いていきたいと思います。」友人A（アメリカ人・科学研究者）

「中国に比べて、空気がきれいだと感じます。また、賃金も高い。なので、私はこの国で働くことができて嬉しいです。」友人B（中国人・エンジニア）

以上のインタビュー結果から明らかになったのは、少子高齢化の中で高度外国人材への需要が高まっているが、いまのところ、彼らには魅力的な国として映っている、ということである。当面心配はいらなさそうである。

満口 倩（遺伝子組換え作物推進センター研究員）

② 記事Aにない情報を集めるために、他のwebサイトの情報を見ましたか？

1. はい
2. いいえ

③ **【この問題は②で「はい」と答えた人にのみお聞きします。】**

記事Aにない情報を集めるために、閲覧したり検索したWebサイトのタイトルや検索に用いたキーワードを下の表の左側に**すべて**書いてください。

また、それらのサイトに記事の主張に関する具体的な記述があった場合、それを表の右側に書いてください。

＊PCでの検索履歴を参照して、記述していただいてもかまいません。

＊先程の問①で，ここで記述したサイトに関係する箇所に印をつけていたか，忘れずに確認してください。

閲覧したり検索をした Webサイト/検索キーワード	記事の主張に関する内容（記事と同じ主張、記事と反対の主張のどちらでもかまいません。）
例）アメリカから見たニッポン！！	このサイトでは、日本は外国人にとって非常に魅力的であると書かれていた。
例）フランスから見たニッポン！！	このサイトでは、記事の主張とは異なり、日本は外国人にとって魅力的でないと書かれていた。
例）Google検索「日本　魅力」	なし

図1　パフォーマンス課題の回答用紙（記事A用）

表3　本研究の市民的オンライン推論能力を評価するルーブリック

観点＼レベル	4	3	2	1
a. 情報発信者の同定	記事の執筆者情報に関する箇所に印をつけており、かつ、外部のwebサイトから、記事作成者の情報を集めようとしている（c. 他の情報の検索）	記事の執筆者情報に関する箇所に印をつけており、かつ、その人がどういった人なのかに関する記述がある	記事の執筆者情報に関する箇所に印をつけているが、その人がどういった人なのかに関する記述がない	記事の執筆者情報に関する箇所に印をつけていない
b. エビデンスの質の評価	記事の証拠や根拠に関する箇所に印をつけており、かつ、外部のwebサイトから、記事上の証拠や根拠の真偽を確かめている（c. 他の情報の検索）	記事の証拠や根拠に関する箇所に印をつけており、かつ、記事の証拠や根拠の質の悪さに関する適切な記述がある	記事の証拠や根拠に関する箇所に印をつけているが、それに関する記述がないもしくは、記述があっても内容が適切でない	記事の証拠や根拠に関する箇所に印をつけていない

3. 調査方法

3.1. 調査対象者および調査手続き

調査は、関西にある学力上位層の私立大学において、2017年度に開講された「教育社会学」「現代の教育」という2つの授業の一環として、コンピュータ室で行われた。調査対象者は上記いずれかの科目を履修した学生計98名である。このうち、記入漏れなどのない90名分（男性58名、女性32名）のデータが分析に用いられた。学年は、学部1年生51名、2年生17名、3年生12名、4年生以上10名（うち、大学院生1名を含む）であった。また、学部系統は、文学部系72名、社会学部系14名、その他4名（法学系、国際関係系）であった。

データの収集・使用にあたっては、授業担当教員から許可を得た上で、調査対象者には研究目的のために使用することを説明し、全員から同意を得た。また、調査の結果は成績評価には反映されないことを調査対象者に伝えた。

調査対象者には、まず調査対象者の特徴（性別、SNSの使用時間など）を知るため、背景質問紙への回答を依頼した。その後、コンピュータ画面上にオンライン記事Aが表示されるようにした。なお、この記事が架空のものであることは調査対象者には伝えなかった。このことを伝えてしまうと、情報発信者が架空の人物であると判断してしまう学生がいると考えたためである。オンライン記事Aを読んだ後で、書かれている主張の真偽を判断させた。真偽の判

断を行う際には、「他の人に聞く以外であればどのような方法を取っても構いません」という教示を行い、自由に書き込みを行えるように白紙のメモ用紙を配付した。次にオンライン記事 B を与え、同様の作業を行わせた。その後、回答用紙を配付し、記入するように指示した。

3.2. 学生による自己評価・相互評価

上記の作業終了後、回答を学生に採点してもらうため、調査対象者にルーブリックの載った採点用紙を渡した。ルーブリックの使い方を説明するため、記事 A 用の回答の具体例を用いて、10 分ほどかけて採点方法を解説した。記事 B については同様の採点方法で行うとのみ伝えた。まず自己評価を行わせ、その後、回答用紙を近くの人と互いに交換させて、相互評価を行わせた。なお、この結果は成績評価には反映されないことを調査対象者に伝えた。最後に、回答用紙および採点用紙を回収した。

3.3. 課題作成者による評価

回答の採点は、大学院で開講される教育評価の授業に参加した第二、第三、第四筆者を含む、課題作成に関わった学生 4 名（大学院生 3 名、学部生 1 名）であった。この 4 名は、授業内で市民的オンライン推論能力に関して時間をかけて議論して理解を深めてきたため、本研究ではこの評価結果を、市民的オンライン推論能力のより妥当な指標[2]とみなし、本稿では「課題作成者による評価」と呼ぶこととする。

課題作成者による評価は、調査対象者による評価とは別途、独立に行われた。まず、評価者 4 名が 2 名ずつに別れ、回答用紙の山を 2 つに分けた。次に、1 枚の回答用紙に対し 2 名が独立して、表 3 のルーブリックを用いて採点を行った。採点結果を用いて算出された κ 係数は .73 ～ .91 であり、Landis & Koch（1977）の基準に基づくと、いずれもかなり高い一致であり、課題作成者間では高い評価者間信頼性を得ることができたといえよう。一致が見られなかった回答に対しては、採点には直接関与していない第一筆者を交えて議論し、一つの得点を与えた。これを持って、最終的な課題作成者による評価結果とした。

3.4. 分析方法

分析においては、①自己評価、②相互評価、の結果を課題作成者による評価結果と照らし合わせて、その一致度を、先行研究に倣い相関係数を用いて分析した（e.g., Campell et al., 2001；Lin et al., 2001；Rudy et al., 2001；Patri, 2002；斎藤他, 2016）。ただし、相関係数が高いことは必ずしも課題作成者による評価結果と学習者による自己評価及び相互評価の結果が近いことを意味するわけではない（Ryan et al., 2007）という指摘に基づき、本研究では全体の中で評価結果が一致した割合（一致率）もあわせて報告する。

4. 結果

4.1. 基礎統計量

背景質問紙により調査対象者の SNS の利用状況に関して次のことがわかった。まず、調査に参加した学生のうち、1 日の中で SNS を見ることがないと回答したものはおらず、調査対象者全員が毎日 SNS を使用していた。そのうち約 8 割は、1 時間以上を費やしていた。また、調査対象者が SNS で流れてくるニュースを多く目にしていた。これを、SNS の利用状況の点で総務省（2017）の調査結果と比較すると、今回の調査対象者は国内の同年代と同様の傾向を持つことが確かめられた。

次に、記事 A、B における観点 a、b の得点の平均値、標準偏差を学生の自己評価、学生の相互評価、課題作成者の評価結果別に算出した（表 4）。それぞれの平均値を見ると、観点 a では、記事 A、B ともに相互評価の採点結果が最も高かった。観点 b では、記事 A、B ともに課題作成者による評価の結果が最も高かった。また、信頼性（内的整合性）の検討のため、課題作成者の評価結果を用いて α 係数を算出した。その結果、観点 a では .63、観点 b では .58 であった。観点 b においては一般的に許容される基準である .60 をやや下回っているが、それほど大きく下回っていなかったため、観点 a、b ともに内的整合性については問題ないと判断した。

表 4　基礎統計量

記事	観点	自己評価	相互評価	課題作成者
A	a	1.09（0.32）	1.11（0.41）	1.09（0.32）
	b	3.01（0.74）	3.08（0.67）	3.21（0.64）
B	a	1.21（0.55）	1.30（0.76）	1.21（0.59）
	b	3.07（0.76）	3.20（0.71）	3.44（0.62）

* 平均値（標準偏差）

図 2、図 3 に各観点における得点の分布を示した。図 2 から明らかなのは、記事 A、B ともに観点 a に関してはほとんどの調査対象者が 1 点にとどまり、記事を読む際に記事執筆者に着目して読むことができていなかったことである。調査後に実施したコメントペーパーにおいても「*記事の執筆者を検索すれば今回の課題の場合だと 1 発で真偽がわかったはずなのに、自分は気づけなかったので、検索して情報の背景を明らかにすることは大切だと思った。*」「*内容*

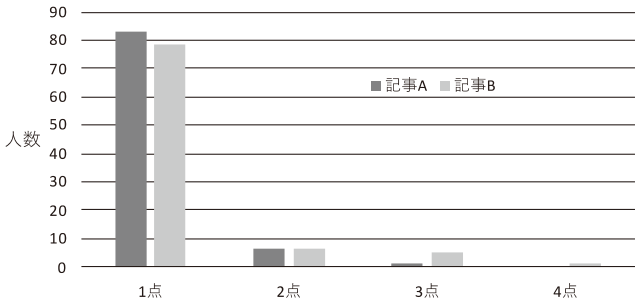


図 2　観点 a の得点分布（課題作成者による評価）

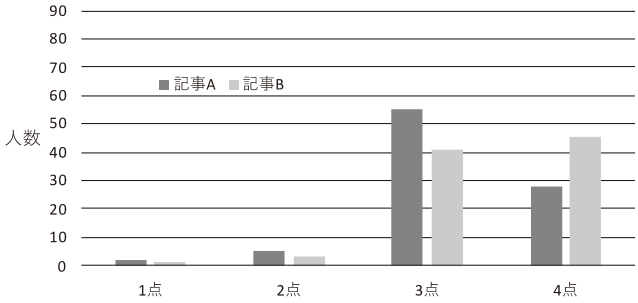


図 3　観点 b の得点分布（課題作成者による評価）

ばかりに気を取られて、誰が書いているかということを気にしていなかった」という主旨のコメントが目立った。

また、図 3 より、観点 b に関しては、記事 A ではその記事のみを注意深く（close reading）読む学生の方が、外部の情報を検索して、その情報から確証しようとする学生よりも多数であったが、記事 B ではラテラルな読みをする学生の方が多くなっていた。観点 b に関することで、「インターネットを使って調べるとき、1 つのサイトだけではなくて、色々な角度から探して調べることが大切だと思った。また、これは本当に正しいのかな?と疑うことも正しい情報を得るために大切なことであるとわかった。」というコメントも見られた。これらの学生によるコメントは、今回のパフォーマンス課題が市民的オンライン推論能力を表出させることができるものであったために得られたものであろう。ここからも開発したパフォーマンス課題の妥当性が確認できる。

4.2. 課題作成者による評価と自己・相互評価の一致度

続いて、学生による自己評価および相互評価の、課題作成者による評価結果との一致率を算出した。表 5 は、各記事で観点ごとにその一致率を示している。例えば、課題 A の観点 a に関しては、課題作成者による評価結果と自己評価結果は 96.7%が完全に一致していた。最も低かったのは記事 B の観点 b に関する課題作成者と自己評価の結果の一致度であるが、それでも 71.1%であった。

次に、各観点での得点の平均を算出し、スピアマンの順位相関係数を算出した。表 6 より、課題作成者による評価と自己評価・相互評価の順位相関係数はいずれも 5%水準で有意であり、観点 a で .90 程度、観点 b で .70 程度であった。

表 5　課題作成者による評価結果との一致率（%）

記事	観点	学生の自己評価	学生の相互評価
A	a	96.7%	95.6%
	b	83.3%	85.6%
B	a	95.6%	93.3%
	b	71.1%	76.7%

表 6　課題作成者による評価結果との相関係数

観点	学生の自己評価	学生の相互評価
a	.92**	.88**
b	.67**	.69**

** $p < .01$

5. 考察

5.1. 学生の評価は妥当な評価に近づきうるか

表 4 のそれぞれの平均値を比較してみると、学生たちが自身あるいは交換した学生の答案を過剰に厳しく、あるいは甘く評価したような痕跡は見られず、課題作成者による評価結果とほとんど差がなかった。これは、先行研究（e.g., Magin, 2001；斎藤他, 2016；Topping, 1998）で指摘されてきた報告内容とは異なっている。また、表 5 の一致率の値を見ると、最も低い場合でも 70%以上、高いものでは 90%以上の絶対的な一致率を確認できた。Reddy & Andrade（2010）では、ルーブリックを用いて採点する際に、一般的に評価者間で 70%以上の一致率が目指されるべきだとしており、この基準を本研究は満たすものであることがわかる。そして、相関係数の値も、表 1 の先行研究での値のように大きなばらつきの見られるものではなく、大きな値を得られた。では、なぜ本研究では、学習者による自己評価及び相互評価の結果が課題作成者による評価結果との一致度（一致率と相関係数）がこのように高くなったのであろうか。この要因を以下では考察する。

第一に、SHEG（2016）のアイデアを援用した、市民的オンライン推論能力のパフォーマンス課題・ルーブリックのデザインにその要因が求められよう。本研究で用いた課題では、学習者が注目した箇所を回答用紙上に再現するように問うことで可視化し、かつ、ルーブリックの各観点でレベルを判断する際に見るべき箇所が明示的となるように開発した。すなわち、客観的に判断可能な箇所を課題に組み込み、

ルーブリックとの対応関係を明確化した。例えば、観点 a「情報発信者の同定」を評価する際に、レベル 1 とレベル 2 以上を分ける「記事執筆者に着目したか否か」の判断は、図 1 の回答用紙の問①で、左下の執筆者情報の周辺に下線部などの印がついているかどうかを見るだけで判断できる。また、他の情報を検索したかどうかも、図 1 の問②への回答を参照すれば容易に分かる。一方で、斎藤他（2016）や木下他（2016）の課題では、回答は、すべてまとまった文章で書かれる。そのため、どこか一定の箇所を確認すれば評価できるわけではない。それを踏まえると、本研究では、ルーブリックにおけるレベルを判断する際に、客観的に判断可能な箇所を組み込んだ課題およびそれと対応したルーブリックのデザインが、一致率や相関係数を高めた一つの理由と考えられる。

第二に、本研究で評価対象となった能力である市民的オンライン推論能力自体の特徴も関係していると考えられる。表 1 で引用した先行研究では、ビジネスプレゼンテーション能力（Campbell et al., 2001）、インタビュー能力（Rudy et al., 2001）、プレゼンテーション能力（Patri, 2002）、アカデミック・ライティング能力（斎藤他, 2016）といった能力が対象となっている。これらは、本研究で扱った市民的オンライン推論能力よりも抽象度が高いため、それらの解釈の幅はより広く、評価者ごとに異なる判断基準を持っていた可能性が高い。このように評価対象とした能力の特徴が、評価結果の一致度に影響していることも十分に考えられる。

第三に、ルーブリックを使っての採点のトレーニングを行ったことが挙げられる。ただし、こちらに関しては 10 分と短く、この影響だけで今回の良好な結果が得られたとは考えにくい。

以上より、課題・ルーブリックのデザイン、評価対象とする能力、採点トレーニングによって、学生による評価結果を、教員や課題作成者が行う、より妥当な評価結果に近づけることが十分可能であることを、本研究の結果は示唆していると言えよう。

5.2. 本研究の意義と限界

これまで国内においては、自己評価や相互評価に関しては、教員による評価と大きくずれるために、その採点結果が妥当であるかどうかに関しては否定的に言及されることが多かった（例えば、木下他, 2016；斎藤他, 2016）。こうしたこれまでの研究に対して、本研究は、客観的に判断可能な箇所を課題に組み込み、ルーブリックとの対応関係を明確化することで採点しやすさを追究した課題・ルーブリックをデザインし、これにより学生による自己評価や相互評価が妥当な評価に近づきうるかどうかを検討した。その結果、学生による自己評価や相互評価と課題作成者による評価結果の一致度は高く、これとほぼ代替可能であることを提示できた。特に、海外における先行研究と比較しても高い一致を示した点は、本研究における極めて有用な知見だといえよう。

一方で、限界も挙げられる。第一に、本研究は単一の大学の一部の学生における調査および分析に留まるため、他の学生集団においても同様の結果が得られるかどうか定かではない。特に、今回調査を実施した対象は、学力の高い集団であったため、評価能力自体も高かった可能性がある。他の集団、例えば低学力の学生が多く所属する集団での市民的オンライン推論能力の評価においても、同様の結果を得られるかどうかに関しては言及できない。ただし、本研究で作成したルーブリックは、他の先行研究（e.g., Oakleaf, 2009）と比べて、採点が容易であると考えられる。なぜなら、1）記述語に程度を表す副詞（たとえば、「十分に」や「適切に」）などを使っておらず、2）調査実施時には各レベルで想定される回答例も提示したためである。ルーブリックでは、記述語にどのような言葉を使用するかが最も難しい側面の 1 つであるとされる（Reddy & Andrade, 2010）が、本研究で用いたルーブリック（表 3）では、評価者が自分で記述語を解釈して判断を下さなければならない余地が少ない。そのため、今回の調査参加者よりも学力的に低い集団においても、高い一致度を期待できる。

第二に、今回得られたデータは、特に観点 a においてレベル 1 の得点に偏っていたために、評価が容易であり、高い一致率を示した可能性を否定できない。レベル 2 以上の該当者の割合がもっと多く、より大きなばらつきが見られる場合に、同様の高い一致度が得られるかは定かではない。この点をより詳細に検討するためには、例えば市民的オンライン推論能力に関しての授業を行ったクラスにおいて、しばらく時間を置いたのちに、この課題を課すという方法が考えられる。この場合、1 点から 4 点まで、回答により大きなばらつきが生まれると想定できるため、1 点に得点が集まったことの影響を、今回の結果との比較を通して検討することが可能となる。

第三に、今回の課題は、成績に関係するような総括的評価のために実施したハイステイクスなテストではない。ハイステイクスなテストであれば、学生が良い成績をとるため、あるいは相互評価で相手に気を遣うためにより高めに採点するであろうことは容易に想像できる。そのため、今回の結果がハイステイクスなテストにおいても適用可能かどうかは、本研究の結果からは言及できない。

第四に、課題の一部に妥当性への脅威があった可能性がある。具体的には、課題 A と課題 B における得点の分布を比較すると、観点 b において、課題 A から課題 B の

間で、4 点の該当者が 15 名増えていることがわかる。同様の大きな分布の変化は、観点 a においては観察されなかった。この要因として課題自体における訓練効果、環境や手続きの影響が考えられる。例えば、課題 A に取り組む最中に、周りの学生がパソコンで調べ物をしていることに気づいて、課題 B においてこれを実行した可能性が考えられる。ただし、いかなる要因がこのような得点分布の変化を起こしたかは本研究では特定できない。このような妥当性への脅威の可能性は、事前の段階では全く予期されなかったため、今後これを排除できるような方法を模索する必要があろう。

5.3. 今後の課題

学生による評価結果が妥当な評価に近づきうる場合、パフォーマンス評価の導入における障壁は非常に低くなると考えられる。市民的オンライン推論能力は現代の情報化社会の中でまさに市民として必要な能力である。これを評価するにはパフォーマンス評価を用いるのが適切であるが、必ずしも教員が評価する必要はなく、学生の評価結果をそのまま授業に活かすという方向性もあることを本研究の知見は示している。例えば、図 2 の課題作成者による評価結果と同等のものが学習者自身の評価結果によって得られたとしよう。この評価結果に基づけば、指導者は、情報発信者（記事執筆者）に注目して記事を読ませること、そして記事内にあるエビデンスの質を、記事外にある他の情報をもとに判断することを強調した指導をすれば良いこととなる。こうしたビジョンを実現していくために、今後の課題として以下のことが検討されるべきである。

第一に、総括的評価としての利用の可能性を検討することである。今回の研究により、診断的評価・形成的評価として学習者自身の評価結果を活用できる可能性が見出された。今後は、総括的評価として利用すると学生に教示した場合（すなわちハイステイクスなテスト）においても、本研究と同様の結果が得られるかどうかを検討することが課題である。

第二に、今回用いたような、客観的に判断可能な箇所の組み込みやルーブリックとの対応関係の明示は、どのような能力を対象とした場合、どのような課題において適用可能なのかを検討する必要がある。例えば、パフォーマンス課題としてしばしば取り上げられるレポート課題においても、ルーブリックとの対応関係の明示が部分的に適用できる可能性がある。具体的には、レポート課題において、序論―本論―結論で構成することをルールとして定めれば、それぞれの箇所で求める記述があるかどうかの判断（序論で問題提起がなされているか、本論で主張の根拠が述べられているか、など）は、対応関係が比較的明確であるために行いやすい可能性がある。ただしその場合、レポートライティングをかなり定型化することになり、課題の質を損ねることへ繋がりかねない。そのため、パフォーマンス課題とルーブリックの観点の対応関係の設け方には注意を要するであろう。一方で、「論理性」「文章の流れの美しさ」といった、全体で判断しなければならないルーブリックの観点に関しては、このように対応関係を明示することが困難と考えられる。このように、今回用いた方法の適用範囲の探索に加えて、SHEG（2016）で用いられた正誤問題の組み込みや本研究で用いたような、印の有無の確認といったような方法以外に、どのような評価方法が考えられるかも今後研究していくことでパフォーマンス評価の現場での普及に寄与しうるだろう。

第三に、学生はどのような場合において、より妥当な評価を行うことができるのかがより多面的に検討されるべきである。藤原他（2007）は、相互評価の交換方法に着目して、お互いに交換して評価するのではなく、自分が評価する答案の持ち主は自分以外の答案を採点する方式のほうが、より妥当な評価になりやすいことを指摘している。また、学習者自身の理解度が高いほど、より妥当な評価を行うことができる傾向があるという研究もある（木下他, 2016）。

こうした実証的な知見の蓄積が、学習者自身が妥当な評価を行うことができるための条件を明らかにする。その結果として、教員による評価が困難な状況において、学生を巻き込んでパフォーマンス評価を行い、得られた評価結果をもとに指導を改善するという方向性がより現実味を帯びてくる。そして同時に、専門家の判断がどうしても必要な場面と学生による評価でも参考にできる場面を識別していくことも期待できる。

本研究で試みたように、パフォーマンス評価の実行可能性を高めることは、その得られた評価結果を指導に活かし、学生の知識・技能の活用能力を育成していくために非常に重要である。本研究はこうした方向性へとつなげていくための一つの試みである。今後同様に、パフォーマンス評価の実行可能性に着目した研究が蓄積されることで、「評価」を中心とした教育実践デザインの普及・拡充が期待される。

付記

本研究は長沼祥太郎・杉山芳生・澁川幸加・浅川裕子, Jeong Hanmo・土岐智賀子・山田勉・松下佳代（2018）「実行可能性に考慮したデジタル・リテラシー評価の開発」『第 24 回大学教育研究フォーラム発表論文集』207. を発展させたものである。

注

[1] 結論の正しさ自体は、問題の問④（本稿では記載なし）で問うているが、今回の評価の対象としなかった。なぜなら、記事 A、B ともに、結論の正しさの検証自体が学術的に非常に難しいためである。そのため、今回は結論を導くプロセスに焦点を当てている。

[2] 自己評価や相互評価の文脈では、教員の評価に近いことをもって、「妥当である」とされる場合がある（cf. Ashenafi, 2017）。一方で、心理測定論における妥当性とは、「目的とする構成概念を実際に測れているかどうか」を指し（cf 村山, 2012）、教員による評価自体も、妥当性を満たしているかの検討対象となる。これらを踏まえ、本研究では、教員による評価と同様、課題作成者による評価は、学生自身による評価に比べると、より妥当な評価であるとの仮定に立つ。その上で、学生による評価が課題作成者の評価に近いことを、「妥当な評価に近い」と表現する。

引用文献

Ashenafi, M. M. (2017). Peer-assessment in higher education-twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education*, *42*(2), 226–251.

Bouzidi, L., & Jaillet, A. (2009). Can online peer assessment be trusted? *Educational Technology & Society*, *12*(4), 257–268.

Campbell, K. S., Mothersbaugh, D. L., Brammer, C., & Taylor, T. (2001). Peer versus self-assessment of oral business presentation performance. *Business Communication Quarterly*, *64*(3), 23–40.

中央教育審議会（2008）.『学士課程教育の構築に向けて（答申）』（http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2008/12/26/1217067_001.pdf）（2019 年 8 月 23 日）

中央教育審議会（2012）.『新たな未来を築くための大学教育の質的転換に向けて―生涯学び続け、主体的に考える力を育成する大学へ―（答申）』（http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2012/10/04/1325048_1.pdf）（2019 年 8 月 23 日）

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

藤原康宏・大西　仁・加藤　浩（2007）.「公平な相互評価のための評価支援システムの開発と評価：学習成果物を相互評価する場合に評価者の選択で生じる『お互い様効果』」『日本教育工学会論文誌』31（2）, 125–134.

布施　泉・岡部成玄（2010）.「多段階相互評価法による学習の実践と効果」『日本教育工学会論文誌』33（3）, 287–298.

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.

木下　涼・藤原康宏・永岡慶三（2016）.「共通レポートを用いた相互評価における他者評価の正確性と理解度との関係」『日本教育工学会論文誌』40（supple.）, 217–220.

国立大学図書館協会（2015）.『高等教育のための情報リテラシー基準』（https://www.janul.jp/j/projects/sftl/sftl201503b.pdf）（2019 年 8 月 23 日）

河野昭彦・斉藤博嗣・佐々木大輔・平澤一樹・須田達・鶴谷奈津子（2014）.「学生の相互評価によるアクティブラーニング型授業」『工学教育』62（6）, 62–67.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement of categorical data. *Biometric*, *33*, 159–174.

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*(4), 420–432.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, *26*(1), 53–63.

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類に基づいて―」『京都大学高等教育研究』18, 75–114.

松下佳代（2016）.「共通教育における学習成果の直接評価―成果と課題―」『大学教育学会誌』38（1）, 29–34.

松下佳代・畑野　快・斎藤有吾・浅井健介・河合道雄・周　　静・田中正之・丁愛美・Nikan Sadehvandi・蒲　雲菲・星野俊樹・松井桃子・長沼祥太郎（2015）.「ルーブリックの意義と課題―ルーブリックの批判的検討を踏まえて―」『第 21 回大学教育研究フォーラム発表論文集』196–197.

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: Civic reasoning in a social media environment. *American Educator*, *41*(3), 4–9.

文部科学省（2017）.『平成 27 年度の大学における教育内容等の改革状況について（概要）』（http://www.mext.go.jp/a_menu/koutou/daigaku/04052801/__icsFiles/afieldfile/2017/12/13/1398426_1.pdf）（2019 年 8 月 23 日）

村山　航（2012）.「妥当性概念の歴史的変遷と心理測定学的観点からの考察」『教育心理学年報』51(0), 118–130.

生田目康子（2004）.「ピア・レビューを伴うグループ学習の評価―一斉型プログラミング授業への適用―」『情報処理学会論文誌』45（9）, 2226–2235.

西片　裕（2017）.「学生によるルーブリックの作成と自己評価が自律的動機付けに与える影響」『日本教育工学会論文誌』41（supple.）, 69–72.

西岡加名恵・石井英真・田中耕治（編）（2015）.『新しい教育評価入門』有斐閣コンパクト.

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interater reliability. *Journal of the American Society for Information Science and Technology*, *60*(5), 969–983.

尾澤重知・望月俊男・江木啓訓・國藤　進（2004）.「グループ間相互評価による協調学習の再吟味支援の効果」『日本教育工学会論文誌』28（4）, 281–294.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19*(2), 109–131.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448.

Rudy, D. W., Fejfar, M. C., Griffith, C. H., & Wilson, J. F. (2001). Self- and peer assessment in a first-year communication and interviewing course. *Evaluation & the Health Professions*, *24*(4), 436–445.

Ryan, G. J., Marshall, L. L., Porter, K., & Jia, H. (2007). Peer, professor and self-evaluation of class participation. *Active Learning in Higher Education*, *8*(1), 49–61.

斎藤有吾・小野和宏・松下佳代（2016）.「パフォーマンス評価における教員の評価と学生の自己評価・学生調査との関連」『日本教育工学会論文誌』40（Supple.）, 157–160.

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, 1–4.

SHEG (Stanford History Education Group) (2016). *Evaluating information: The cornerstone of civic online reasoning*. (https://stacks.stanford.edu/file/druid:fv751yt5934/SHEG%20Evaluating%20Information%20Online.pdf)（2019 年 8 月 19 日）

総務省（2017）.『情報通信白書平成 29 年版 PDF 版』（http://www.soumu.go.jp/johotsusintokei/whitepaper/h29.html）（2019 年 8 月 23 日）

田中耕治・水原克敏・三石初雄・西岡加名恵（2011）.『新しい時代の教育課程』有斐閣アルマ.

Topping, K. (1998). Peer assessment between students in college and universities. *Review of Educational Research*, *68*(3), 249–276.

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』26, 189–205.

山内祐平（2003）.『デジタル社会のリテラシー』岩波書店.

Research Papers

Can Students' Self-Assessment or Peer-Assessment in Performance Assessment be Close to Valid Assessment? A Case from the Field of Civic Online Reasoning

Shotaro Naganuma[1], Yoshiki Sugiyama[2], Sachika Shibukawa[2], Yuko Asakawa[3], and Kayo Matsushita[4]

([1]University Education Innovation Initiative, Kyushu University, [2]Graduate school of Education, Kyoto University,
[3]Japan International Cooperation Agency (JICA), [4]Center for the Promotion of Excellence in Higher Education, Kyoto University)

The purpose of this study is to reveal whether learners' assessment results in performance assessment can be close to valid assessment. In a case from the field of civic online reasoning, the authors first developed performance tasks and a rubric. Referring to previous research, our performance tasks sought the facilitation of scoring by incorporating objectively judgeable points and clarifying the relationship to the rubric. We had 90 private university students engage in performance tasks and conduct self- and peer-assessments using the rubric. Agreement between self- and peer assessment results from the students and those of the test developers themselves were calculated with two indicators: concordance percentage and correlation coefficient. The results indicated that high agreement was observed between students' self- and peer assessments and test developers' assessments. Our study revealed that when using certain assessment methods for targeted competencies, assessment results by learners themselves can be nearly equivalent to those of experts. This finding suggests that learners' self- or peer assessment results can be utilized to improve the feasibility of performance assessment.

Keywords: performance assessment, feasibility, civic online reasoning, self-assessment, peer assessment