大学教育学会誌　第42巻　第１号　2020年５月

〈事例研究論文〉

協議ワークを取り入れたピアレビューによる学生の自己評価力向上の効果検証

岩　田　貴　帆
（京都大学大学院教育学研究科）

The Effect of the Peer Review with Discussion for the Moderation of Self- and Peer-Assessment on Improving Students' Self-Assessment Ability

Takaho Iwata,
(Graduate School of Education, Kyoto University)

In performance evaluation, students' ability to assess one's own performance is crucial for effective learning. However, university students tend not to assess their own performance appropriately. Peer review with discussion for the moderation of self- and peer-assessment could improve students' self-assessment ability. To examine the effectiveness of the peer review, the researcher conducted a case study with 87 university students. An assessment exercise using rubrics was introduced and then a peer review was conducted. The exercise could contribute to understanding all aspects and levels of the rubrics. The statistical analysis showed that students' self-assessment ability was significantly improved through the peer review. Also, students significantly improved their performance after the peer review. It is therefore suggested that the assessment exercise is a prerequisite for effectively conducting the peer review.

〔**キーワード**：自己評価力，ピアレビュー，学習としての評価，パフォーマンス評価，評価練習〕

１．問題と目的

⑴　パフォーマンス評価における自己評価力の重要性

大学教育において重要性が高まっているパフォーマンス評価は，課題に取り組む学生自身が評価主体となり，評価それ自体が学びとなる「学習としての評価」という意義ももつ（松下，2016）．「学習としての評価」とは，Earl（2013）が教室で行われる評価をその主体と目的から整理した３つの分類である「学習の評価」「学習のための評価」「学習としての評価」のうちの１つである．「学習としての評価」は，学生が評価主体となり学習を自分でモニターし，調整するために行う評価とされている．そこでは，学生の自己評価が重要であり，学生は自分が何をわかっていないのかを認識し，次に何をすべきかを決定するプロセスによって学習を前進させていくという．このように学生がパフォーマンスの遂行→評価→改善というプロセスに取り組むことで，より質の高いパフォーマンスを生み出すことができ，その課題で発揮することが求められる能力の獲得が促されると考えられる．

パフォーマンスを自己評価し，改善するプロセスにおいて，その自己評価は適切でなければならない．なぜなら，学生の自己評価が的外れなものならば，その後に適切にパフォーマンスの改善に取り組むことができないからである．もしも学生が現状のパフォーマンスを過大に自己評価してしまう場合には，改善は不要であると判断するだろう．逆に過少に自己評価してしまう場合には，既にすぐれたパフォーマンスができているにも関わらず，不必要な改善に時間をかけるだろう．したがって学生が適切に自己評価できる力が必要といえる．

学生が適切に自己評価することは必要である一方，容易ではない．知識の活用や問題解決等を求めるようなパフォーマンス課題を評価するという行為には，Sadler

(2010) が指摘するように，「評価基準」「質」「タスクコンプライアンス」といった概念の理解や暗黙知といった専門性を要するため，学生が適切に評価することは難しいと考えられる．実際，国内の大学におけるレポート課題において，ルーブリックを用いた学生の自己評価と教員評価はズレがあり，相関はほぼなかったことが報告されている（斎藤ら，2017a）．よって，複数の評価者間の採点のズレを小さくするためのツールであるはずのルーブリックであっても，学生がルーブリックの意味を理解して使いこなすことは容易ではないと考えられる．

仮に，学生が容易に使いこなせるようにとルーブリックの記述語をパフォーマンスの表面的な特徴（e.g.「誤字脱字が2つ以下」）にすると，そのパフォーマンス課題を通して評価すべき能力を十分に評価できない，妥当性の低い評価基準となる恐れがある．さらに，学生がパフォーマンスの表面的な特徴の要件を満たすことで満足してしまい，高めるべき能力を用いずにパフォーマンスを遂行してしまう恐れすらある．

したがって，教員は妥当性の高い評価基準を学生に共有した上で，その評価基準を用いて学生が適切に自己評価できるよう，何らかの手立てを講じる必要がある．そうすることで，学生は自らのパフォーマンスを適切に評価し，改善することが可能になり，そのプロセスを通して，課題が要求する能力を高めてゆけると考えられる．

以上を踏まえて，学生が適切に自己評価する力の向上に焦点を当てた研究には，大学教育に貢献しうる意義があると考えられる．

⑵ 自己評価力の定義

本研究で焦点を当てる，学生が適切に自己評価する力，すなわち自己評価力を定義するにあたりEvaluative Judgement概念（Tai et al, 2018）や，ダニング＝クルーガー効果（Kruger & Dunning, 1999）などの自己評価に伴うバイアスを考慮して定義する．

Evaluative Judgementとは，「自分や他者の作品・活動（work）の質について決定を下す能力」（Tai et al., 2018, p.471）と定義されている概念であり，「質は何によって構成されているかを理解すること」「自分や他人の作品・活動を評価するにあたってその理解を適用すること」の2つの構成要素からなるとされている．Panadero et al.（2019）によれば，Evaluative Judgementを有する学生は，質・評価基準・スタンダードについて理解しており，その理解に基づいて目標を設定したり，自分のパフォーマンスをモニタリングできるため自己調整学習（Zimmerman, 2000）を促進する効果を持つ．このようなEvaluative Judgementは，本研究で焦点を当てる，学生が自分のパフォーマンスを適切に評価することで適切に改善できるようになることを目指す自己評価力と大きく重なるものである．そこで自己評価力の定義にあたりEvaluative Judgementを元にする．

ただし，Evaluative Judgementは，評価対象を自分の作品・活動に限定せず，他者の作品・活動を評価する場合も定義に含めているため本研究でそのまま使用することはできない．なぜなら，Kruger & Dunning (1999) によれば，自己評価には様々な認知バイアス（動機，利己，記憶の選択的想起，他者の有能さの無視など）の影響を受け，能力の高い人ほど自分の能力を過小評価し，能力の低い人ほど自分の能力を過大評価する傾向（ダニング＝クルーガー効果）があるからである．Boud & Falchikov（1989）は自己評価に関する11の量的研究をレビューした結果，同様の傾向が広く存在することを明らかにしている．こういった自己評価に伴う認知バイアスが存在するため，他者を適切に評価できることと，自分を適切に評価できることは区別して考える必要がある．そこで本研究では，評価対象を自分のパフォーマンスに限定して自己評価力を定義する．

以上を踏まえ，本研究では自己評価力の定義を「評価基準を理解し，それを自分のパフォーマンスに適用することで，目標とのギャップを適切に把握する力」とする．

⑶ ピアレビューによる自己評価力向上とその抑制要因

自己評価力の向上に焦点を当てた先行研究には，教員からのフィードバックに着目した研究も存在する（e.g. 斎藤ら，2017b）が，本研究では教員の評価負担がより小さいピアレビューに着目する．一般にピアレビューとは，学生同士でパフォーマンスを交換し，評価し，評価結果を本人にフィードバックし合う教授法である．本研究では，他の学生のパフォーマンスの質を判断する行為をピアレビューにおける評価とし，その評価結果を本人に伝える行為をフィードバックとする．パフォーマンスを見せ合い，評価し，フィードバックし合う一連の活動がピアレビューにあたる．

ピアレビューでは，ピアのパフォーマンスを評価するプロセスにおいて，自己評価力が高まることが明らかにされている（Nicol et al., 2014）．一方，ピアからフィードバックを受ける際に，無批判にその内容を受容してしまう場合，むしろ自己評価力が高まらないと考えられる．例えば，Yucel et al.（2014）の実証研究では，ピアレビュー前には，自分のレポートは改善の余地があるということを適切に認識できていたにも関わらず，「君のレポートはとてもよくできているよ」というピアからの評価を信じた結果，自分のレポートの改善は不要であると判断してしまった学生の事例が報告されている

(pp.978–979)．すなわち，ピアからの評価を無批判に受け入れてしまう場合，ピアレビューによる自己評価力を高める効果が抑制されると考えられる．したがって，ピアから評価を受けるプロセスに何らかの工夫を設けたピアレビューの実施方法が必要である．

⑷ 協議ワークを取り入れたピアレビュー

岩田・田口（2020）は，ピアからの評価を本人が無批判に受け入れてしまう課題を解決しうる，協議ワークを取り入れたピアレビュー[1)]を開発している．その実施手順は，プレ自己評価（ステップ１）→作品を交換してピアと評価し合う（ステップ２）→協議ワーク（ステップ３）→ポスト自己評価（ステップ４）である（図１）．最も肝要な機能を担う協議ワーク（ステップ３）とは，プレ自己評価とピアからの評価の得点やその根拠を比較し，ズレが見られる箇所についてズレた理由を学生同士で話し合う学習活動である．プレ自己評価とピアからの評価を比較することによって，ピアからの評価は必ずしも正しいわけではないことを前提としつつ，より適切に自分のパフォーマンスを評価するための参考情報としてピアからの評価を受け取ることができると考えられる．

岩田・田口（2020）では開発に加え，大学授業での実施と分析も行ってはいるものの，協議ワークを取り入れたピアレビューの前後で自己評価力の有意な向上は確認できていない．その理由の１つには，評価基準として用いたルーブリックの全観点・全水準の記述語を学生が理解できるような学習活動をしていない状態で協議ワークを取り入れたピアレビューに取り組んだことが考えられる．ルーブリックの各水準が示す，パフォーマンスの質的な差異を理解するためには，異なる水準に該当するパフォーマンスのサンプルの比較を通してその質的な差異を確かめることが重要と考えられる．

そこで本研究では，ルーブリックの全観点・全水準の記述語の理解を促すような学習活動を実施した上で，協議ワークを取り入れたピアレビューの前後で学生の自己評価力が向上するかの効果検証を行い，実践的示唆を得ることを目的とする．

２．方法

⑴ 実践のフィールド

効果検証のため，大学の授業における実践を通して得られたデータを分析する．実践のフィールドは京都大学の教養科目「社会学Ⅰ」(2019年度前期）である．筆者は本フィールドにTAとして関わっており，授業の設計や進行，学期末レポートやその草稿の採点補助に関わった．

本授業科目は，同内容で授業曜日が異なる２つのクラスが存在する．両クラスのシラバスや授業内容は完全に同じで，実際の授業の進行にも大きな差は生じていないことから，本研究では両クラスの受講者全体を対象とする．受講登録者数は木曜クラス54名，金曜クラス58名であった．授業の成績は，授業参加40点，学期末レポートの草稿10点，学期末レポート50点の配点であった．

授業の学習到達目標は「社会学的思考法を用いて，現代のさまざまな社会現象や自分自身の人生・生活の背景にある『しくみ』（社会構造とコミュニケーションの相互作用）を，基礎的な水準で分析し説明できるようになること」である．授業担当者による講義資料や講義中の説明によると，社会学的思考法とは，コミュニケーションと構造の影響関係に着目する思考方法のことである．ここでいうコミュニケーション，構造は社会学の用語であり，これらの影響関係に着目しながら社会現象の分析・説明を行うことによって，その社会現象の暗黙の前提を指摘したり，それを踏まえた有効な問題解決策を提示したりすることができるという．この学習到達目標の達成度を評価するパフォーマンス課題として次のような

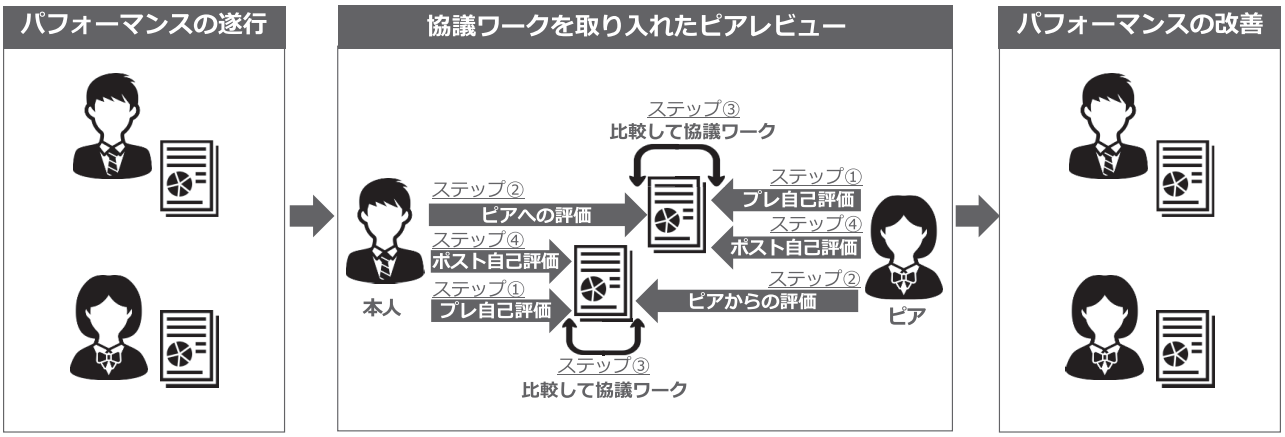

図1　協議ワークを取り入れたピアレビューの実施手順

表1 「社会学Ⅰ」のレポート課題評価用ルーブリック

	観点1　社会現象	観点2　問題	観点3　解決策	観点4　客観性
レベル4	レベル3の水準に加え，参考書において暗黙の前提となっている箇所に切り込んで批判・補足している．	レベル3の水準に加え，問題を示すことによって，参考書における暗黙の前提を相対化するような批判・補足をしている．	レベル3の水準に加え，解決策は実現可能性が高いことを示せている．	レベル3の水準に加え，自分の主張に対して予想される反論にも応じながら，議論を展開できている．
レベル3	着目した社会現象について参考書における自分の論述に必要な箇所の要旨を説明しており，その理解に誤りが全くない．	問題を示すにあたり，社会学的思考法を用いることによって，問題とする社会現象の背後にある構造を，説明できている．	解決策を示すにあたり，社会学的思考法を用いることによって，「問題」で説明した構造に変化をもたらしうる解決策を，示している．	データや事例を適切に示しながら議論を展開することで，説得力を持っている．
レベル2	着目した社会現象について，参考書における要旨を説明しているが，自分の論述に必要な箇所を説明しきれていない．または，その理解に一部誤りがある．	問題を示すにあたり，社会学的思考法を用いてはいるが，説明している構造は問題とする社会現象とは関係の薄いものである．	解決策を示すにあたり，社会学的思考法を用いてはいるが，問題とする社会現象の背後にある構造に変化をもたらしうる解決策は，示せていない．	データや事例を示してはいるものの，適切でない箇所がみられ，説得力を持つに至っていない．
レベル1	着目した社会現象について，参考書における要旨を説明しているが，自分の論述に必要な箇所を説明しきれていない．かつ，理解に一部誤りがある．	問題を示すにあたり，社会学的思考法を用いておらず，問題とする社会現象の背後にある構造を，説明できていない．	解決策を示すにあたり，社会学的思考法を用いておらず，問題とする社会現象の背後にある構造に変化をもたらしうる解決策は，示せていない．	データや事例を示しておらず，自分の考えを述べるに留まっている．
レベル0	着目した社会現象について，参考書における自分の論述に関係のある箇所を全く説明していない．または，致命的な理解の誤りがある．	そもそも問題とする社会現象を示していない．	そもそも解決策を示していない．	データや事例を示しておらず，自分の考えも述べていない．

学期末レポートを設計した．

指定文献で展開されている論を批判もしくは補足するため，扱われている社会現象のうち1つに着目し，社会学的思考法（コミュニケーションと構造の影響関係に着目する）を用いて，その社会現象に関連する「問題」と「解決策」を客観的に示しなさい（1,000〜2,000字程度）．

このレポート課題の評価基準として表1のような4観点5水準（レベル0〜4）のルーブリックを用いた．このルーブリックは，評価者がレポートの質を適切に判断できる評価基準となること，特に水準間が等間隔な尺度となることを意図して次の手順で作成したものである．まず，授業担当者と筆者が，課題指示と仮ルーブリックを作成した．次に，筆者と協力者（過去に当該授業担当者の授業を履修した者）の2名が作成した複数のサンプルレポートを，授業担当者と筆者が仮ルーブリックを用いて独立に採点した．採点結果を比較し両者で協議を行い，仮ルーブリックの記述語を修正し，合意に至った．加えて，本フィールドの1年前に実施された同科目において同じルーブリックを用いて実際の学生のレポートの採点し，妥当な評価基準になっていると判断した．

田中（2008）は，評価基準の各尺度に具体的な記述語が書かれたルーブリックは，対応する典型的なサンプルと対応付けながら複数の者が合議によって作成した場合には，「その尺度はたんなる『名義尺度』ではなく『順序尺度』さらには『間隔尺度』をめざすものと言えよう」（p.143）と述べている．これを踏まえ，本ルーブリックは間隔尺度を目指したものといえるが，統計学的手続きによって水準間の等間隔性を担保したものではない点において，厳密には間隔尺度ではない点に留意を要する．

⑵ 評価練習の実施方法

本研究の目的に対応して，評価基準として使用するルーブリックの全観点・全水準の記述語の意味を学生が理解することを促すような評価練習を実施する．評価練習とは，パフォーマンス課題の評価基準を理解するために，教員が用意したパフォーマンスのサンプルを学生が評価基準を用いて評価した上で，教員による評価やその根拠を解説するという学習活動（Rust et al., 2003; Tai et al., 2018）を指す．

本実践では，以下のような方法で評価練習を実施することにより，学生がルーブリックの全観点・全水準を理解することを意図した．まず，サンプルレポートとして，前年度の受講生が提出した期末レポートの中から，表2に示すような教員評価が付与されたレポートを選出した．選出の際，4種類のレポートによって各観点の全ての水準（1〜4）が網羅されるようにした．ただし，社会現象の観点についてはレベル1のレポートを選出できなかったため，レベル3が2つ存在する．

これらのサンプルレポートを，第5回の授業（木曜ク

表2　評価練習で用いたサンプルレポートの教員評価

	社会現象	問題	解決策	客観性
サンプルレポートA	3	3	3	1
サンプルレポートB	4	4	4	3
サンプルレポートC	2	1	1	4
サンプルレポートD	3	2	2	2

ラス：2019年５月16日，金曜クラス：５月17日）において学生にルーブリックと共に配布した．次に，学生の座席順により自分がどのレポートを担当するのかを決定し，担当レポートを読んで評価の得点と根拠をワークシートに記入するよう学生に指示した．その後，それぞれレポートA～Dの評価を担当した４人でグループを構成し，評価を共有しあい，修正する必要がある箇所は修正することを目指してディスカッションを実施した．A～Dの教員評価は各観点の中でレベルが異なっていることはディスカッション前に伝えた．ディスカッション終了後，教員から各レポートの教員評価と根拠を解説した．以上の所要時間は，進行の説明やワークシート等の配布・回収を含め約70分であった．

このように評価練習を実施することで，授業の時間的制約の中で学生は異なる水準に該当するレポートを比較しながらその質的な差異を確かめることができ，ルーブリックの各水準の記述語の理解を深めることを意図した．

⑶ 協議ワークを取り入れたピアレビューの実施方法

第10回の授業（木曜クラス：６月27日，金曜クラス：６月28日）で協議ワークを取り入れたピアレビューを実施した．学生が持参したレポートの草稿に対して，プレ自己評価（ステップ１）→作品を交換してピアと評価し合う（ステップ２）→協議ワーク（ステップ３）→ポスト自己評価（ステップ４）の手順で実施した．ルーブリックを記載したワークシートを用いて，プレ自己評価，ポスト自己評価の得点やその根拠などを記入するよう指示した．ポスト自己評価のワークシートでは，「改善点・提出までにすること」を記入する欄を設けた．ワークシート記載内容は一切成績に影響しないことを学生に伝えた．授業後，草稿レポートとワークシートを回収した．以上の所要時間は，進行の説明やワークシート等の配布・回収を含め約70分であった．

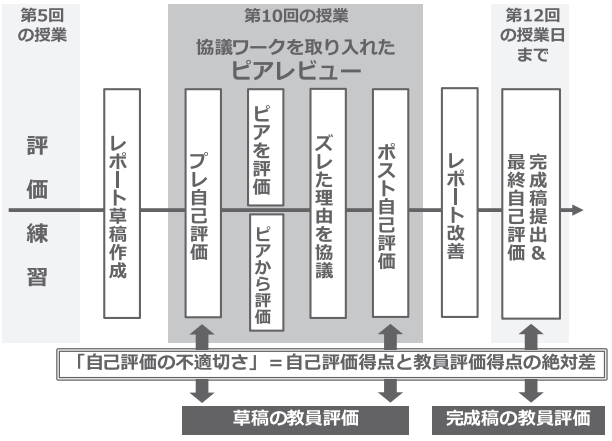

図2　授業の展開と分析に使用するデータ

学期末レポート提出期限は，木曜クラス：７月11日，金曜クラス：７月12日であった．それまでの２週間，学生は草稿レポートの内容を必要に応じて改善した上で，完成稿として学期末レポートを提出した．その際，完成稿に対する自己評価得点も併せて提出するよう指示した．なお，草稿提出から完成稿提出期限までに実施された第11回・第12回の授業では，レポート課題の「指定文献」の内容（子育て支援政策）に関係する動画視聴とグループディスカッションを実施した．

⑷ 分析に使用するデータと自己評価力の指標

以上のような授業展開により得られた，草稿レポートに対する教員評価得点と，ピアレビュー前後それぞれの自己評価得点，加えて，完成稿レポートに対する教員評価得点と自己評価得点を本研究の分析に用いる（図２）．

教員評価は，TA（筆者）が下採点として得点とその根拠を記録した後，授業担当者が記録を参考にしながら本採点したものである．意見が一致しない箇所は協議・合意した．下採点と本採点の一致率は94.5％であった．

自己評価力を量的に捉えるには，レポートの適切な評価結果を示す基準値が必要である．先行研究（斎藤ら，2017b）の議論を参考に，最も適切な評価結果と考えられる，ルーブリックを用いた教員評価得点を基準値と位置付ける．そこで，自己評価得点と教員評価得点の絶対差を「自己評価の不適切さ」（岩田・田口，2020）とする指標を用いる．すなわち，プレ自己評価とポスト自己評価については草稿の教員評価との絶対差を算出し，完成稿の自己評価については，完成稿の教員評価との絶対値を算出する（図２）．この自己評価の不適切さが小さい値を示すほど，適切に自己評価できていること，すなわち自己評価力が高いことを意味する．

なお先述のように，本フィールドで使用したルーブリックは統計学的に厳密には間隔尺度ではなく順序尺度である．順序尺度として扱うならば，自己評価得点と教員評価得点の差を算出することはできない．しかしながら，本研究の目的である，特定の教授法の前後で学生の自己評価力が向上するかの効果検証に際しては「自己評価と教員評価がどの程度離れているのか」を扱うことに意味がある．したがって本研究では本フィールドのルーブリックを間隔尺度として扱い，自己評価と教員評価の絶対差を「自己評価の不適切さ」と指標化するが，統計学的厳密性の点で限界を有することに留意を要する．

⑸ 研究目的に対応する分析手法

まず，協議ワークを取り入れたピアレビューが学生の自己評価力の向上を促す効果があるかを検証するにあたり，協議ワークを取り入れたピアレビューの前後で自己評価の不適切さがどのように変化するかを統計的に分析

する．さらに，実践的示唆を得るため，自己評価の対象であるレポートのバージョンが草稿から完成稿へと変化した場合に自己評価の不適切さがどのように変化するのかも考察の対象とする．そこで，プレ自己評価，ポスト自己評価，完成稿の自己評価という３回の「自己評価の不適切さ」の変化を分析する．また，ルーブリックの４つの観点ごとに変容の仕方に違いがある可能性も考えられるため，ルーブリックの合計得点ではなく各観点の自己評価の不適切さを分析に用いる．よって，「自己評価の不適切さ」を従属変数，「タイミング（プレ，ポスト，完成稿）」と「観点（ルーブリックの観点１～４）」をそれぞれ参加者内要因とする二要因分散分析を行う．

加えて，実践的示唆を得るため，協議ワークを取り入れたピアレビューの後に学生がパフォーマンスを適切に改善できているかどうかも検討する．そこで，「教員評価」を従属変数，「バージョン（草稿，完成稿）」と「観点（ルーブリックの観点１～４）」をそれぞれ参加者内要因とする二要因分散分析を行う．

それぞれの分散分析において主効果に有意差が見られた要因については，平均値差の有意性を多重比較（Shaffer法；入戸野，2004）する．統計分析ソフトは，IBM SPSS Statistics ver.25を使用する．ただし，分散分析後の多重比較における効果量算出やShaffer法の機能がSPSSにはないため，HAD ver. 16.056（清水，2016）を用いた．有意水準は全て $\alpha = .05$とする．

3．結果と考察

⑴ 要約統計量

研究協力の同意が得られた学生のうち，ピアレビューに出席し，各ワークシートを提出した87名を分析対象とした．表３に，自己評価・教員評価・ピアからの評価の平均得点と標準偏差SD，指標である自己評価の不適切さの平均値と標準偏差SDを示した．図３に，各観点の自己評価の不適切さの推移を表記し，変化の効果量として標準化平均値差d（Hedges，1981）をグラフ内に示した．表４に，教員評価１～４点に対する自己評価・ピ

表3　要約統計量（自己評価・教員評価および自己評価の不適切さ）

上段：平均 下段：SD	観点１	観点２	観点３	観点４	観点 合計		観点１	観点２	観点３	観点４	観点 合計
プレ 自己評価	2.80	2.74	2.76	2.81	11.11	プレ 自己評価の 不適切さ	0.49	0.64	1.09	0.59	2.82
	0.68	0.65	0.88	0.78	1.79		0.53	0.76	0.99	0.67	1.64
ポスト 自己評価	2.97	2.85	2.75	2.83	11.40	ポスト 自己評価の 不適切さ	0.36	0.61	1.03	0.51	2.51
	0.66	0.66	0.78	0.78	1.60		0.48	0.72	0.88	0.65	1.68
完成稿の 自己評価	3.31	3.33	3.32	3.36	13.32	完成稿の 自己評価の 不適切さ	0.38	0.78	1.30	0.57	3.04
	0.51	0.50	0.70	0.73	1.59		0.49	0.88	1.06	0.62	1.92
草稿の 教員評価	3.00	2.49	1.99	2.70	10.18	草稿の ピアからの 評価	3.25	3.02	3.01	3.05	12.32
	0.55	0.86	1.03	0.79	2.04		0.63	0.75	0.86	0.69	1.63
完成稿の 教員評価	3.09	2.62	2.22	2.94	10.87						
	0.50	0.85	1.10	0.74	2.00						

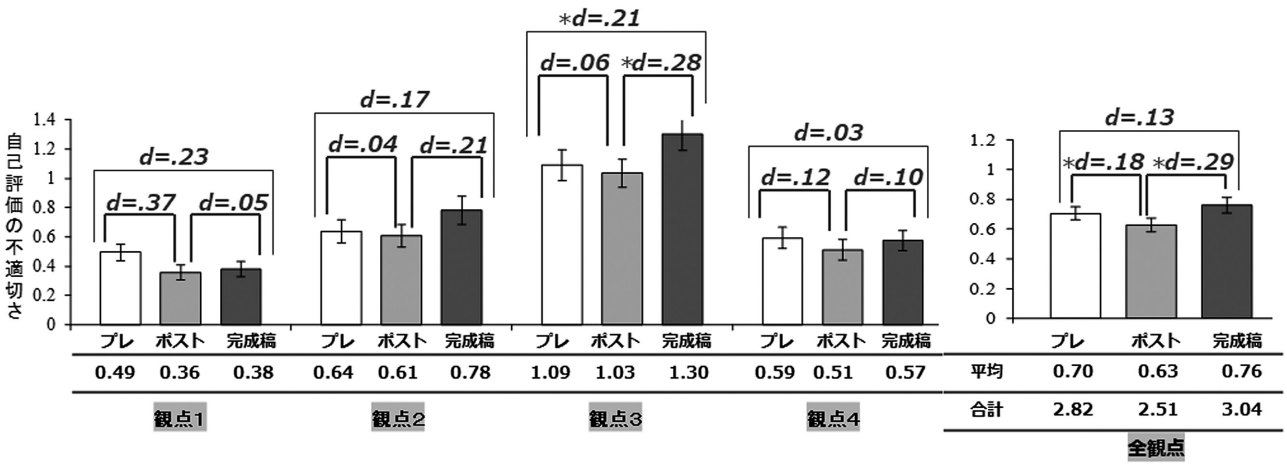

図3　自己評価の不適切さの推移と効果量

表4　教員評価ごとの自己評価・ピアからの評価

	教員評価	1点	2点	3点	4点
草稿	N	61	67	185	34
	プレ自己評価	2.53	2.65	2.88	3.00
	ピアからの評価	2.74	2.83	3.22	3.55
	ポスト自己評価	2.49	2.63	2.97	3.35
完成稿	N	50	50	196	52
	完成稿自己評価	3.25	3.12	3.30	3.71

表5　レポート改善群と非改善群の自己評価力の推移

	観点1		観点2		観点3		観点4	
	N	変化量	N	変化量	N	変化量	N	変化量
改善群	7	+0.28	11	−0.72	16	−0.13	16	−0.31
非改善群	80	−0.00	76	+0.30	71	+0.35	71	+0.14
		n.s.		＊＊		†		＊

変化量＝「完成稿の自己評価の不適切さ」−「ポスト自己評価の不適切さ」
†p＜.10　＊p＜.05　＊＊p＜.01

アからの評価の平均を，全観点をまとめて示した．

⑵　自己評価の不適切さの変化に関する分散分析の結果

「自己評価の不適切さ」を従属変数，「タイミング」と「ルーブリックの観点」を参加者内要因とする，二要因分散分析を実施した．その結果，タイミングの主効果（F（2，172）＝4.760，p＝.011，偏η^2＝.52），観点の主効果（F(3，258)＝26.400，p＝.000，偏η^2＝.235）と，ともに統計的有意な影響があった．交互作用は有意傾向であった（F(6，516)＝2.165，p＝.060，偏η^2＝.025）．交互作用が有意である場合には，要因ごとの単純主効果の検定を行うこととなるが，有意傾向に留まることから，ここでは単純主効果の検定は行わないことを判断した．

タイミングに関する多重比較の結果，プレと比較してポストの自己評価の不適切さが有意に小さくなった（調整p＝.040）．これは，自己評価力が協議ワークを取り入れたピアレビューの前後で高まったことを意味する．一方，ポストと比較して完成稿の自己評価の不適切さが有意に大きくなった（調整p＝.011）．これは，協議ワークを取り入れたピアレビューの直後ではレポートの草稿を適切に自己評価できるようになったものの，レポートの完成稿では適切に自己評価できなかったことを意味する．プレと完成稿には有意な差が見られなかった（p＝.242）．

観点に関する多重比較の結果，自己評価の不適切さは観点1よりも観点4が有意に大きく（調整p＝.032），観点4と観点2には有意差がなく（p＝.139），観点2よりも観点3が有意に大きい（p＜.001）ことがわかった．これは，観点1は学生にとって理解・適用が比較的容易な評価基準であった一方，観点3は逆に比較的困難であることを意味する．

⑶　教員評価の変化に関する分散分析の結果

「教員評価」を従属変数，「バージョン（草稿・完成稿の2水準）」と「ルーブリックの観点（観点1〜観点4の4水準）」を参加者内要因とする，二要因分散分析を実施した．その結果，バージョンの主効果（F（1，86）＝31.598，p＜.001，偏η^2＝.269），観点の主効果（F(3，258)＝27.793，p＜.001，偏η^2＝.244）と，ともに統計的有意な影響があった．交互作用は非有意であった（F(3，258)＝2.074，p＝.114，偏η^2＝.024）．

観点に関する多重比較の結果，教員評価は観点3よりも観点2が有意に高く（調整p＝.005），観点2よりも観点4が有意に高く（調整p＝.015），観点4よりも観点1が有意に高かった（調整p＝.009）．

以上の結果から，協議ワークを取り入れたピアレビューの後に学生が草稿レポートを修正した結果，教員評価は有意に向上しており，適切な改善ができていることが明らかとなった．また，交互作用の検定からその向上の仕方については観点による有意な違いは見られなかった．

さらに，草稿から完成稿にかけてパフォーマンスを改善できた群と，改善できていない群に群分けし，自己評価の不適切さの変化量に両群で差があるかを対応なしt検定で調べた（表5）．正の値は，自己評価力の低下を意味し，負の値は，自己評価力の向上を意味する．観点2・3・4の改善群は，自己評価力が向上しているのに対して，非改善群は，自己評価力が低下しており，その差は統計的有意または有意傾向）である．観点1については逆の傾向がみられるものの，そもそも改善できた学生が7人と少ないこともあり，統計的有意な差ではない．

⑷　考察

まず，プレ自己評価と比較して，ポスト自己評価の不適切さが統計的有意に減少したことについて考察する．この2回の自己評価の間に学生が経験したのは協議ワークを取り入れたピアレビューのみであることから，本教授法の効果として自己評価力が向上したといえよう．自己評価力の向上は，表5の教員評価と自己評価の相関からも見て取れる．パフォーマンスの低い学生ほど過大に自己評価し，パフォーマンスの高い学生ほど過少に自己評価する傾向（ダニング＝クルーガー効果）がプレ・ポストの両方で現れているが，その傾向はポストの方が小さい．このことから本教授法によって，自己評価に伴うバイアスが小さくなった可能性が示唆される．

この結果に影響を与えた可能性のある要因として，本研究の目的から特に重要と考えられる，ピアからの評価

と，評価練習の２点について検討する．

１点目に，協議ワークを取り入れたピアレビューとは，ピアからの評価を本人が無批判に受け入れてしまう課題を解決することを意図した教授法であった．表４の教員評価１～２点では，ピアからの評価は，プレ自己評価よりもさらに過大評価をしていた傾向がわかる．もしも，ピアからの評価を本人が無批判に受け入れたならば，ポスト自己評価はさらに甘くなるところだが，実際には教員評価に近づいている．また，表４の教員評価３～４点では，プレ自己評価とピアからの評価の中間あたりにポスト自己評価が変化しており，それは教員評価に近づく方向である．これらを総合すると，学生はピアからの評価を無批判に受容することなく参考情報として活用した結果として，自己評価が適切になった可能性が示唆される．

２点目に，本実践の評価練習では，ルーブリックの全観点・全水準を網羅するようなサンプルレポートを用いてグループディスカッションを行うことで，学生がルーブリックの各水準の評価基準について理解を深めることを意図した．このことが影響し，協議ワークを取り入れたピアレビューで自己評価とピアからの評価のズレた理由を話し合う際に，評価基準の理解に基づいた議論ができたのではないだろうか．今回の研究デザインでは，評価練習そのものの効果を検証することはできなかったが，岩田・田口（2020）と比較すると，ルーブリックの全観点・全水準の理解を促すことが，協議ワークを取り入れたピアレビューが効果を発揮する前提として重要であることが示唆される．

次に，観点ごとの違いについて検討を行う．観点１～４のいずれにおいても，自己評価の不適切さの平均値はプレからポストにかけて減少している．効果量dを観点ごとに比較すると，$d = .37$である観点１に対して，他の３つの観点は$d = .04$～$.12$と小さい．観点１は，学生にとって理解・適用が比較的容易な評価基準であったという多重比較の結果と併せて考えると，評価基準の理解・適用が比較的容易な観点ほど，協議ワークを取り入れたピアレビューの効果が発揮されやすい可能性が示唆される．逆に，観点３は学生にとって理解・適用が比較的困難な評価基準であった．その理由として，問題とその解決策を示すことを求めるレポート課題の内容を反映したルーブリックを作成した結果，観点３（問題の解決策）は観点２（問題の発見）を前提としている箇所があり，より複雑な評価基準であったことが考えられる．このような評価基準においては，自己評価とピアからの評価の得点や根拠を比較したり，ズレた理由について学生同士で話し合ったりしても，ピアからの評価を参考に自分のレポートを見直すことに十分に繋がらない可能性が考えられる．このことからも，協議ワークを取り入れたピアレビューの前提として，学生が評価基準を十分に理解することを促す学習活動を実施しておく必要があるといえよう．これに対応する実践的示唆としては，理解・適用が難しいと予想されるルーブリック観点に関しては，予め評価練習を豊富に実施しておく等が考えられる．

また，草稿におけるポスト自己評価と比較して完成稿の自己評価の不適切さが有意に大きくなった点について考察する．表４から，完成稿の教員評価が１点や２点の場合，過大に自己評価する傾向が草稿よりもさらに大きくなっていることがわかる．つまり，自己評価の不適切さの平均値が大きくなったのは，低いパフォーマンスの学生の自己評価が特に影響している．さらに表５が示すように，レポートを改善できた群は，自己評価の不適切さが低下する傾向があり，レポートを改善できなかった群は，自己評価の不適切さが上がる傾向があった．すなわち，教員評価が１や２のケースにおいて，学生はレポートを改善できたと考えて完成稿を提出し，自己評価を高くつけたが，教員からすればルーブリック得点が向上するほどの質的な改善に至っていないというケースが多数存在したということであろう．特にパフォーマンスの低い学生に対して，修正したパフォーマンスが評価基準を満たすほどの改善になっているかを自分で判断できるようになるための，さらなる支援が必要といえよう．

最後に，教員評価の変化について考察を行う．協議ワークを取り入れたピアレビューの後に学生が草稿レポートを修正した結果，教員評価は有意に向上したことがわかった．また，その向上の仕方については観点による有意な違いはなく，どの観点も教員評価が向上していた．本研究のデータでは，草稿から完成稿にかけてのレポートの質向上に関する因果関係を特定することはできないが，影響を与えた可能性があるのは，この間に学生が経験した協議ワークを取り入れたピアレビューと第11・12回の授業内容の２点である．１点目について，協議ワークを取り入れたピアレビューの最後に学生がワークシートの「改善点・提出までにすること」欄に記入した内容を筆者が見たところ，実際に各学生が修正を試みた点に概ね合致していた．このことから，学生は協議ワークを取り入れたピアレビューがレポートの質向上に影響を与えたと考えられる．２点目について，授業内容はレポート課題の指定文献の内容に関係するが，レポートで要求するような社会学的思考法の活用に焦点化した授業ではなかった．このことから，授業内容の影響は限定的と考えられ，主に協議ワークを取り入れたピアレビューがレポートの質向上に繋がった可能性が示唆さ

れる.

以上のように，本研究のフィールドでは，評価練習ならびに協議ワークを取り入れたピアレビューを実施し，概ね期待通りの効果を確認できた．ただし，本フィールドは学習到達目標・パフォーマンス課題・評価基準が相互に対応するように設計されているという特徴をもつ．特に，評価基準であるルーブリックは，具体的なレポートと対応付けながら複数名で記述語を検討し，作成したものであった．それゆえ，抽象度の高い記述語にも裏付けとなるパフォーマンスが存在しており，評価練習においては，全観点・全水準を網羅するようなサンプルレポートを活用しながら学生は評価基準の理解を深めていくことができたと考えられる．したがって今回の結果の過度な一般化は避けるべきであり，少なくとも，学習到達目標・パフォーマンス課題・評価基準が十分に対応している状況が必要と考えられる.

4．まとめと今後の課題

本研究では，協議ワークを取り入れたピアレビュー（岩田・田口，2020）の実践を通して，その前後で学生の自己評価力が向上していることを効果検証した．本研究から得られた知見は，以下の２点にまとめることができる.

まず１点目に，パフォーマンス課題における自己評価力を高めるための教授法として，協議ワークを取り入れたピアレビューの有効性を実証的に検証できた点である．学生同士でパフォーマンスを評価・フィードバックし合うのみの従来のピアレビューでは，ピアからの評価を無批判に受け入れてしまう場合，自己評価力が向上しないという課題があった（Yucel et al., 2014）．自己評価とピアからの評価を比較して，得点やその根拠のズレが生じた理由を学生同士で話し合う協議ワークを取り入れることで，この課題が解決されうると考えられる．本研究では，協議ワークを取り入れたピアレビューの前後で自己評価力が統計的有意に向上することを検証できた．また，協議ワークを取り入れたピアレビューは，パフォーマンスの質向上にも繋がる可能性が示唆された.

２点目に，協議ワークを取り入れたピアレビューが効果を発揮するための前提として，事前に実施する評価練習において評価基準の理解を促すための工夫が重要であるという点である．本研究の実践では，ルーブリックの全観点・全水準に対応したサンプルレポートを用いた評価練習によって，学生が評価基準を一定程度理解した上で，協議ワークを取り入れたピアレビューに取り組むことができたと考えられる.

一方，本研究には，以下３点の課題が存在する.

１点目に，本研究では評価練習の実施方法が間接的に，のちのピアレビュー前後の自己評価力の向上に影響を与えている可能性が示唆されたが，その直接的な効果は検証できていない．今後は，評価練習の実施方法についても先行研究を踏まえた開発と，評価練習そのものが自己評価力に与える直接的な効果の検証を行うことが必要である．特に，パフォーマンスの低い学生を念頭に，自己評価力向上の手立てをさらに検討することも重要だ.

２点目は，効果検証を行ったフィールドの限定性である．本研究では京都大学の学生を対象とし，教養科目の社会学の授業をフィールドとした．様々な要因が影響しているであろう今回の結果について過度な一般化は避けるべきだ．特に，学習到達目標・課題・評価基準を相互に対応づけた課題設計は重要な前提といえる．したがって，他大学や他分野の授業，専門科目やPBL，卒業研究といった場面でも同様の効果を得るためには，どのような支援や工夫が必要なのか，さらなる検討を要する.

３点目は，本研究では自己評価力を量的に捉えて効果検証を行うため，ルーブリックを用いた教員評価を最も適切な評価と位置づけて自己評価と教員評価の絶対差を用いた点である．Sadler（2010）が指摘するように評価には暗黙知や鑑識眼が関わっていることを踏まえると，ルーブリックを用いても教員評価が常に最も適切な評価とは限らない．また，ルーブリックの等間隔性が統計学的に担保されていない点で分析結果には留意を要する．それゆえ，本研究の自己評価力の指標化の方法では，自己評価力を捉えきれていない可能性がある．自己評価力をより精緻に捉える方法の１つとして，自己評価力が，パフォーマンスの質の向上にどのように影響を与えているかを実証的に検討することが考えられる．Panadero et al.（2019）が提示するように，自己評価力が向上する結果としてパフォーマンスの質向上に繋がることが理論的に考えられる．本研究の分析は，協議ワークを取り入れたピアレビューが，自己評価力とパフォーマンスの質向上のそれぞれに与える影響を検討するに留まったが，各学生の自己評価力と教員評価得点の関係についてより精緻な分析を行うことによって，学習を促進する評価活動について実証的な知見を高める余地がある.

大学教育における自己評価力や「学習としての評価」について授業実践を踏まえて実証的に論じた研究は，斎藤ら（2017b）や小野ら（2018）など限られているが，それらも述べるように，この研究領域は今後ますます議論の活発化が予想される．本研究は上述のような課題も有するが，その議論に貢献できれば幸いである.

注

1）岩田・田口（2020）で相互評価活動とされている教授法を，本研究ではピアレビューと表記する．ピアとの相互作用を通して自分のパフォーマンスをレビューする（re view：再び見る）という語義を踏まえた．

付記・謝辞

本研究は筆者の修士論文の一部を発展させたものです．指導教員の田口真奈准教授（京都大学高等教育研究開発推進センター）をはじめ貴重なご指導を頂いた先生方，院生の方々，そして研究にご協力くださった柴田悠准教授（京都大学大学院人間・環境学研究科）と受講生の皆様に深く感謝申し上げます．

文献

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18* (5), 529–549.

Earl, L. M. (2013). *Assessment as learning: using classroom assessment to maximize student learning* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics, 6*, 107–28.

岩田貴帆・田口真奈（2020）「パフォーマンス課題における自己評価力を高めるための協議ワークを取り入れた相互評価活動の開発」『日本教育工学会論文誌』43（Suppl.），173–176.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

松下佳代（2016）「アクティブラーニングをどう評価するか」松下佳代・石井英真編著『アクティブラーニングの評価』東信堂，pp.3–25.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102–122.

入戸野宏（2004）「心理生理学データの分散分析」『生理心理学と精神生理学』22(3)，275–290.

小野和宏・斎藤有吾・松下佳代（2018）「PBLを評価する改良版トリプルジャンプにおける『学習としての評価』の要因」『京都大学高等教育研究』24，35–44.

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self-and co-regulated learning: the role of evaluative judgement. *European Journal of Psychology of Education, 34*(3), 535–557.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28* (2), 147–164.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35* (5), 535–550.

斎藤有吾・小野和宏・松下佳代（2017a）「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（Suppl.），157–160.

斎藤有吾・小野和宏・松下佳代（2017b）「ルーブリックを活用した学生と教員の評価のズレに関する学生の振り返りの分析：PBLのパフォーマンス評価における学生の自己評価の変容に焦点を当てて」『大学教育学会誌』39(2)，48–57.

清水裕士（2016）「フリーの統計分析ソフトHAD：機能の紹介と統計学習・教育，研究実践における利用方法の提案」『メディア・情報・コミュニケーション研究』1，59–73.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481.

田中耕治（2008）『教育評価』，岩波書店.

Yucel, R., F. Bird, J. Young, & T. Blanksby. (2014). The road to self-assessment: Exemplar marking before peer review develops first-year students' capacity to judge the quality of a scientific report. *Assessment & Evaluation in Higher Education, 39* (8), 971–986.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In Boekaerts, M., Pintrich, P. R., & Zeidner, M. (Eds.) *Handbook of self-regulation*(pp.13–39). San Diego, CA, Academic Press.