

Japanese-Language Article

「学習を促す評価 (Learning-Oriented Assessment (LOA))」としての自己評価とピア評価:日本の高校生のライティング授業を対象として

Self-Assessment and Peer Assessment as Learning-Oriented Assessment: A Focus on the Writing Classes of Japanese Senior High School Students

大井洋子 (おおい ようこ)

Yoko Suganuma Oi

清泉女子大学

Seisen University

自己評価やピア評価のような学習者による評価は英語教育の現場で活用されてきたが、それぞれの評価タイプと「学習を促す評価(learning-oriented assessment (LOA))」との関係は論じられてこなかった。また、高校生を対象としたライティングの授業は、他の三技能であるリーディング、スピーキング、リスニングに比して、授業の頻度が少なく、その内容も文法等の学習に偏る傾向があり (Mulvey, 2016; 行森, 2018)、ライティング授業の改善が待たれている。本研究はライティング授業を活性化させる方法の一つとして学習者による評価を提案し、その有効性と学習者情意への影響を考察することを目的とする。研究には293人の15～18歳の日本人高校生が参加した。生徒は自己評価グループ(147名)とピア評価グループ(146名)に分かれ、それぞれのタイプの評価訓練を受けた後に10日間のあいだに5

<https://doi.org/10.37546/JALTJJ46.1-4>

JALT Journal, Vol. 46, No. 1, May 2024

回英作文を書き、その直後にそれぞれの評価活動を行った。この集中的なライティング活動の後に、研究に参加した12人の生徒と研究を観察した二人の英語教師が半構造的面接に参加した。自己評価とピア評価は似ているが異なる効果を英語ライティング力の向上や情意にもたらし、「学習を促す評価(LOA)」を効果的にするためには、教師が適切に介入することが必要であることがわかった。

This study explored how student assessments, that is, self-assessment and peer assessment, were related to learning-oriented assessment (LOA) by exploring the reliability and the effects of each assessment type on students' writing performance and learner affect. A total of 293 students aged 15 to 18 years participated in the survey. They were divided into two groups, self-assessment group (147 students) and peer assessment group (146 students). Each group had a rater training and then had five consecutive writing sessions for ten days. Just after writing English compositions, each assessment group evaluated their own writing or a peer's writing. After the sessions, 12 students and two teachers were interviewed. It was found that each assessment type had similar but different effects on writing performance and learner affect. It was also found that teachers were expected to intervene in self- and peer assessment to make learning-oriented assessment more successful.

キーワード: 自己評価;ピア評価;英語教育;学習を促す評価

Keywords: English education; learning-oriented assessment; peer assessment; self-assessment

コミュニケーション力育成を目指した英語教育の改革が叫ばれ、スピーキングやライティングなどの相手あるいは読み手に意図を伝達する「産出的(Productive) 技能」の強化の必要性が強調されている。だが、スピーキングに比してライティングを扱う授業の頻度が他の技能の授業頻度に比べて少ないことが指摘されている(文部科学省, 2018)。また、ライティングの授業内容が文法や語彙学習に大きく時間を割く傾向があることや(文部科学省, 2011)、中高では学年の上昇につれ、ライティング活動がさらに減少し、訳読によるリーディングや語彙・文法指導に時間が多くとられている(Mulvey, 2016; 行森, 2018)。こうした現状の中でライティング授業を充実させていくために、学習者による評価を積極的に取り上げることが提案したい。なぜならば、学習者による評価は、「学習を促す評価」(Learning-oriented assessment (LOA): Turner & Purpura, 2016)として、学習者を学びの中心に据えることが可能なので、ライティング授業の活性化につながると考えられるからだ。学習者による評価には、自分の学習成果を評価する自己評価(Boud, 1992)と学習者同士による評価であるピア評価(相互評価・他己評価・他者評価)がある(Topping, 1998)。こうした学習者による評価は、形成的評価 (formative assessment) の一つでもある (Black 他, 2003; Wiliam 他, 2000)。なぜならば、形成的評価は、学習目標に照らし合わせて学習者がどの程度まで学習目標に到達しているかを評価し、次の学習段階へ導き、学習を促すことを目標にする評価であるからだ (Leahy & Wiliam, 2012)。つまり、学習者が立場を変えて評価者になることによって、学習目標と今の学習レベルとの開きを意識できることが形成的評価の目的である (Leahy & Wiliam, 2012)。教授者が自己評価とピア評価というそれぞれの特質を理解すれば、自己評価とピア評価を効果

的に授業に活用できると考えられる。しかしながら、自己評価とピア評価を比較研究したものは少なく、研究対象の多くは大学生で、特に日本の高校生を対象にした研究は少ない (Oi, 2021)。また、これら二つのタイプの評価法が「学習を促す評価 (LOA)」としてどのような意味があるのかは探究がされていない。なぜならば、自己評価やピア評価の先行研究は、その信頼性に着目して教師評価の代替評価として可能かどうかを検証したものが多く (Oi, 2021)、学習を促す評価としての役割を追求したものは少ないからだ。学習を促す評価では、教師と学習者は評価の責任を共有すべきであるとされている。生徒が評価活動に取り組むことによって、学習を促す評価を前進させる可能性があることからその探求は喫緊の課題である。また、生徒評価を授業に導入することによって、学習者の英語力の進展にどのような影響を与えるかについても、教授者にとっては欠くことのできない命題だと言える。したがって、本研究は、自己評価とピア評価を比較することによって、それぞれの評価法の特徴を信頼性とライティング能力、そして情意に与える影響から分析し、学習を促す評価として期待できることは何かを考察する。

先行研究

先行研究の傾向として、自己評価とピア評価をそれぞれ単独で分析するものが多く、この二つを直接比較して、それぞれの評価方法の特徴を明確にした研究は少ない (Oi, 2021)。また、先行研究のメタ・アナリシスによると、1980～2018年までに行われた学習者による評価に関する研究の約7割はヤング・アダルト (young adult) と呼ばれる大学生を対象としている (Oi, 2021)。学習者評価の信頼性は、学習者の教育的背景に影響を受けるという先行研究結果もあるので (Boud & Falchikov, 1989)、本研究は分析が進んでいないヤング・ラーナー (young learner) である高校生対象に研究を行った。

学習を促す評価 (*Learning-oriented assessment: LOA*)

教室評価に対する理論的枠組みとして、「学習を促す評価 (LOA)」には様々な定義がある。その中でも意義深いのは、教育における評価の役割は学習を促すのが目的だとして、評価の意味を再定義したことにある (Carless, 2007)。つまり、評価を行う目的はあくまでも学習の促進であり、評価は学習者の成長のための足場かけ (scaffolding) として機能すべきというものだ。「学習を促す評価 (LOA)」においては、評価は、学習の根本的な要素であり、教授内容や指導方法を決定する拠り所となるものなのである (Mok, 2012; Ploegh, 2009)。言い換えれば、総括的評価と形成的評価を分離せず、協同的に学習者の学びを促進させ、公式的な評価と教室評価を協調的に機能させようという意図が土台となっている (Carless, 2007; Jones & Saville, 2016)。「学習を促す評価 (LOA)」には、理論的枠組みにとどまらず、この枠組みを基本にして様々なEFLのコンテキスト合わせて言語教育に応用していく流れがある。例えば、言語技術の向上の観点からLOAが論じられたり (Green, 2017; Hamp-Lyons, 2017)、実用性 (Alsowat, 2022) や教師のLOAに対する認知度 (Derakhshan & Ghiasvand, 2022) 等も研究の対象となっている。Hamp-Lyon (2017) は、「学習を促す評価 (LOA)」を教室評価に取り入れるモデルを構成する要素として、第一に、学習に焦点を当てた課題、第二に、学習者による評価、そして第三に、学習者に焦点を当て

たフィードバックの3つを掲げている。このモデルでは、学習者と教授者の協同的関係が求められているのが特徴である。このように「学習を促す評価(LOA)」は、比較的新しい理論的枠組みで、それと実践的試みの関係を探究する研究がまだ少ない。よって、本研究では、Hamp-Lyon によって提唱をされた「学習を促す評価(LOA)」を教室評価に採用するモデルの一つである「学習者による評価」に着目をして、その有効性について分析をすることとした。

学習者による評価の信頼性

学習者による評価の信頼性に関する研究は、教師評価と比較する形式で行われてきた。先行研究によれば、自己評価の教師評価に比した信頼性は、ある程度高い傾向があるとされている。例えば、Andre他(2010)の研究によれば、教師評価との相関関係はある程度高く、特に文法評価においては高い傾向があるという結果が示されている。しかし、Runnels (2014)は、大学生を対象にした研究において、自己評価の信頼性は評価項目の数に左右されることを報告している。また、学習者の英語力の高さや評価の訓練度、年齢や文化的背景が信頼性に影響を与えるとしている (Dieten, 1989; Peirce他, 1993)。一方、ピア評価は、教師評価と正の相関関係があるとする研究結果が多いが(Weaver 他, 2011; Orsmond他, 2000)、自己評価と同様に評価項目の理解度や評価項目の数、文化的背景にその信頼性が影響をされる (Falchikov & Goldfinch, 2000; Matsuno, 2009)。また、ピア評価を行う学習者同士が、人間関係を円滑にしようと正直な評価をしないこともある(Oi, 2021)。両方の評価方法ともに、その信頼性には、評価項目に対する理解や訓練、文化や教育的背景が影響を与えることは共通であるが、信頼性を一貫性と厳格度から分析したものは少ないので本研究はこの点について論じることとした。

学習者による評価のライティングへの学習効果

1980~2018年の間に行われた学習者による評価を対象としたメタ・アナリシスによると、学習効果があったとするのは、自己評価で約2割、ピア評価で約3割とされている(Oi, 2018)。学習者によるライティング能力への学習効果については、主に文法や語彙の習得に対するものが多い(McDonald & Boud, 2003; Lee, 2017; Sadler & Good, 2006)。特に、ピア評価に関する研究では、文法や校正力の向上に寄与したという研究があるが(Oi, 2021)、自己評価を対象とした研究では、これらに着目した研究は少ない。また、二つの評価タイプを比較した場合、ピア評価の方が自己評価よりもライティング能力の向上に貢献をしたという結果を示した研究や(Matsuno, 2009)、友達の評価の方が深刻に受け止めるので教師評価よりも効果的であるとする先行研究もある(Zarei & Mahdavi, 2014; Tsui & Ng, 2000)。自己評価とピア評価の学習者に及ぼす影響の共通点としては、学習者の評価項目への理解が深まり、それが学習効果につながることである(Orsmond他, 2000)。しかし、自己評価とピア評価を直接比較して、どちらがライティング能力の改善により効果的かを示した研究は少ないが、教師評価と同様な使用効果をもっているとする研究 (Farrokhi他, 2012)や、自己評価とピア評価を行う学習者グループの間には評価姿勢に違いが見られると論じる研究もある(Aslanoglu他, 2020)。また、信頼性を対象にした先行研究に比して、ライティン

グ能力の向上に着目した研究が少ないために、どちらの評価タイプがより効果的かを判断するのは難しいように思われるので、この二つのタイプの評価法を直接比較し、それぞれのライティング能力向上への影響を分析するのは意義があると考えられる。さらに、自己評価の文法や校正力の伸長に与える影響に関する研究は少ないので、研究の必要があると言える。

学習者による評価の学習者情意に与える影響

自己評価は、学習目的を学習者に明確に意識をさせるので、自己達成感や学習意欲、そして自信の向上に効果がある (Butler & Lee, 2010; Zarei & Usefli, 2015)。Black 他 (2004)は、学習目的を意識化させることによって、内省する習慣を学習者につけることが可能と述べている。一方、ピア評価については、相互評価による協同作業を通して、他者との交流が生まれ、分析力が身につく、評価をする技術が身に付くという先行研究がある (van Gennip 他, 2010; Saito & Fujita, 2009; Cho 他, 2006)。また、自己効力感を高め、緊張感を減少させ精神的安心感を与える (Cho 他, 2006)。Matsuno (2009)によれば、自己評価とピア評価の共通点としては、両評価タイプともに学習に対する責任感や自律性に良い影響を与えると報告している。しかし、両者の違いを論じた研究は少なく、効果的に授業に取り入れていくには、二つの評価タイプの相違点にも着目していくべきだと考える。

目的と研究課題

本研究は、学習者による評価法である自己評価とピア評価をその信頼性・ライティング能力に与える影響の観点から比べ、それぞれの特徴を明らかにするのが目的である。また、それぞれの評価タイプが「学習を促す評価(LOA)」としてどのように機能できるかを論じるものとする。なお、ここでの信頼性は評価者の一致度や安定性の他に評価者間の厳格度や評価者内の一貫性を含む。したがって、以下の研究課題を掲げる。

1. 教師評価に比して自己評価とピア評価の信頼性はあるのか。
2. 自己評価とピア評価はどのようにライティング能力に影響を及ぼすのか。
3. 「学習を促す評価」として自己評価とピア評価の役割は何か。

分析方法

研究参加者

研究には、15～18歳の293名の日本の高校生(表1)と4名の英語教員(日本人2名、英語母語話者2名)が参加をした(表2)。参加生徒の中には英語母語話者はいなかった。多相ラッシュモデル (many-facet Rasch measurement; MFRM; Eckes, 2015; Linacre, 1994) の分析は、英作文評価の総合得点と分析的評価の得点について別々に行った。また、英作文の評価(教師・自己・ピア評価)は、成績には関与しないことを事前に参加者には伝えた。ピア評価については、英作文を授業内で交換する形式で行い、評価表は書き手に直接返却をもらい、書き手が読み終わった後に英語教員が評価表を回収した。

表1.
実験参加生徒のグループ構成

	自己評価グループ				ピア評価グループ			
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
学年	1	1	2	3	1	1	2	3
科目名	I	I	II	III	I	I	II	III
人数	36	36	36	39	36	36	36	38
合計人数	147				146			

注. I = English Communication I; II = English Communication II; III = English Communication III

表2.
実験参加教師の英語教育経験

	JET A	JET B	NET C	NET D
年齢	40	55	38	50
性別	Male	Female	Female	Male
英語教育経験年数	15	30	7	20

注. JET = Japanese English teacher; NET = native English teacher.

研究手順

表1のように、既存のクラスを自己評価グループとピア評価グループの二つに分け、生徒たちは教師と同じ評価表を使い(付録A)、10日間のライティング活動の直前に評価訓練をした。評価訓練1回とライティング・テストを含むライティング活動5回の時に、15分の間に辞書なしで60〜80語の英作文を書いてもらった。英作文を書いた直後に、自己評価かピア評価のどちらかの評価活動を行った。ピア評価では、1名の生徒のライティングを、別な生徒1名が行った。評価表は、高校教師を対象にした高校生の英作文を評価する際に最も重視をする項目の調査結果をもとに作成した「課題達成度」、「構成と内容の一貫性」、「語彙の適切な使用」、「正確な文法の使用」の4観点の4段階(レベル1〜4)、合計16点満点を利用した(Oi, 2019)。評価活動には10分をとり、英作文課題は、高校生を対象とした英語表現の教科書で最も頻度が高いトピックを選んだ(付録B)。なお、英作文課題の難易度は、英検3級の英作文問題と同程度である(Oi, 2018)。

研究方法

量的研究(研究課題1〜2)においては、多相ラッシュモデル(many-facet Rasch measurement; MFRM; Eckes, 2015; Linacre, 1994)を使い自己評価グループの生徒とピア評価グループの生徒の英作文の総合得点と英作文の分析的評価の得点に関する信頼性を分析した。また、同じく多相ラッシュモデルにより英作文能力に対するそ

れぞれの評価者タイプの影響を分析した。ソフトウェアはMinifac (Linacre, 2021)を用いた。

ラッシュモデルは、人と項目の両方を潜在変数(構成概念)上に位置づけ、受験者の能力と項目の難易度を一元的尺度で分析をする項目反応の確立論的モデルである。受験者能力、項目難易度に加えて、第3の相(facet) として評価の安定度を入れることができ、ロジット (logits) を単位とした同じ尺度で受験者能力、項目難易度、評価者の厳しさを推定することができる。ロジットの値が小さいほど、受験者能力は低く、項目難易度は低く、評価者は甘いことを示す。具体的には、受験者能力と他をそれぞれ独立に並べて推定することが可能である。

分析には、4人の教師と293人の生徒、293人の生徒によって書かれた586の英作文(事前と事後のテストの2回分)なので、2930のデータを使った。しかし、一部の生徒が欠席等のために参加ができなかったので、有効となるデータ数は243人の受検者、6人の評価者、事前・事後の2回のテストから得られた2430のデータとなった。内、自己評価グループは244、ピア評価グループは242、そして教師の評価データ数は1944であった(表3)。本研究では、教師評価、自己評価グループの生徒、そしてピア評価グループの生徒をラッシュモデルを使って比較したが、これら3つの評価タイプを同時に比較した先行研究は乏しい(Farrokih他, 2012)。また、ラッシュモデルを使用時に、評価者グループまたは評価者個人として扱うかは、研究参加者の数や教育環境にも作用される(Aslanoglu他, 2020; Farrokih他, 2012)。本研究では、評価者一人一人の評価結果をラッシュモデルを使って分析を行ったが、紙幅の関係で評価グループとして報告をしている。また、本研究のoccasion相は、pre-testとpost-test が異なっていることから“Occasion by Task”相とした。Pre-testのmeasureの値は .03、一方、post-testのmeasureの値は-.03で、値の差は.06で、先行研究の結果に基づき(Oi, 2018)、二つのタスクの困難度はほぼ同程度とみなした。また、タスクは高校生を対象とした教科書の頻出度上位のタスクから選んだ(Oi, 2018)。

表3.
多相ラッシュモデル分析に使われた有効なデータ数

	参加者数		評価数	欠席者合計	有効な評価数		評価数合計
自己評価グループ	147	147 students x 2 tests	294	25	122	122 students x 2 tests	244
ピア評価グループ	146	146 students x 2 tests	292	25	121	121 students x 2 tests	242
Teachers	4	4 teachers x 2 tests x 293 students	2,344	0	4	4 teachers x 2 tests x 243 students	1,944
テスト回数	2				2		
合計		2,930				2,430	

質的研究(研究課題1～3)においては、12名の生徒(自己評価とピア評価の各グループから英作文の評価結果をもとに6名ずつ抽出、表4)と2名の教員(日本人1名JETB、英語母語話者1名NETC、表2)に対して半構造的面接データを行った。なお、実験前後の英作文の点数は、表2に記された4名による英語教師の評価平均である。生徒には自分の考えを語りやすい日本語で、英語母語話者を含むことから教師には英語で、学習者による評価の信頼性やライティング能力の向上に関する効果についての面接を実施し録音をした(表5)。

表4.
インタビューを受けた生徒一覧

	全体平均	SHSS 1	SHSS 2	SMSS 1	SMSS 2	SLSS 1	SLSS 2	PHSS 1	PHSS 2	PMSS 1	PMSS 2	PLSS 1	PLSS 2
実験前の英作文の点数	12.22	14	15	12	12	10	9	15	15	12	12	9	8
実験後の英作文の点数	12.43	15	14	12	12	10	9	14	14	12	12	11	11
年齢	16.3	15	16	18	16	17	16	18	15	16	17	16	16
性別		男子	女子	男子	女子	男子	女子	男子	女子	男子	女子	男子	女子

注. SHSS = 自己評価グループで高得点生徒(high-scoring students in the self-assessment group); SMSS = 自己評価グループで中程度得点生徒(middle-scoring students in the self-assessment group); SLSS = 自己評価グループで低得点生徒(low-scoring students in the self-assessment group); PHSS = ピア評価グループで高得点生徒(high-scoring students in the peer assessment group); PMSS = ピア評価グループで中程度得点生徒(middle-scoring students in the peer assessment group); PLSS = ピア評価グループで低得点生徒(low-scoring students in the peer assessment group)

表 5.
半構造面接質問項目

Interviewees	
Students	Teachers
- あなたの意見では、生徒評価(自己評価またはピア評価)は英語のライティング能力を向上させるのに効果的ですか。 - もしそうならば、どんな点で重要なのですか。 - 生徒評価(自己評価またはピア評価)は自分の気持ちや学習態度に何か影響を与えましたか。 - 生徒評価(自己評価またはピア評価)の良い点、または悪い点は何ですか。	- In your opinion, is student assessment important for developing writing ability? - If so, in what ways? - Did you observe that learner affect such as anxiety and autonomy did students change by using self-assessment/peer assessment? - How can you utilize self-assessment/peer assessment in writing classes? - What are the effects and challenges of student assessment?

録音は筆者が書き起こしを行い、グラウンデッド・セオリー(Glaser & Strauss, 1967)を用いて分析をした。これは、他の質的分析に比べ、グラウンデッド・セオリーは、人間の語りや行為の意味を分析、コード化し、それら上位の概念的カテゴリーを考察するのに確立された手法だからである。グラウンデッド・セオリーは、データから概念を抽出し、概念同士を関連付けようとする方法(戈木, 2018, p.2)であり、抽象度の異なる4種類の概念、プロパティ(特性)とディメンション(次元)、ラベル、カテゴリーが核となっている。その内、抽象度が低いプロパティとディメンションは分析を通して使う分析の中心的なもので、それには5つの役割があるとされている(戈木, 2018)。第一に、データから概念を抽出する土台となること、第二に概念を理解するヒントを提供すること、第三に概念同士につながりを見出すこと、第四に現象の中にある変化のパターンを提示すること、そして最後に分析者がデータの解釈をする際に解釈の理由を説明できることである。

本研究では、抽出のレベルを増すために半構造インタビューのデータを5段階に分けて分析を行った:(1)データを内容ごとに切片に分け、(2)オープン・コーディング、及びアキシャル・コーディング(軸足コーディング)でカテゴリーを関連付け、(3)カテゴリーを現象ごとにセオレティカル・コードに分類、(4)セオレティカル・コードのカテゴリー関連図の作成、(5)各カテゴリー関連統合図の中心となっているカテゴリーを、プロパティとディメンションで関係づけた。コーディングには、著者の他、日本人英語教師(JETA)がコーダーとして加わった。著者のモデル・コーディングの後に、例を付したガイドラインを作成し、JETAがコーディングを見て矛盾点がないかについてのチェックを行った。分析に不一致があった場合は、その都度話し合いを行いコーディングの修正を行った。

分析の最終段階で、生徒評価と学習を促す評価(LOA)の関係を考察するために探索的に質的・量的研究分野を統合した。

量的研究分析結果

1. 教師評価に比して自己評価とピア評価の信頼性はあるのか。

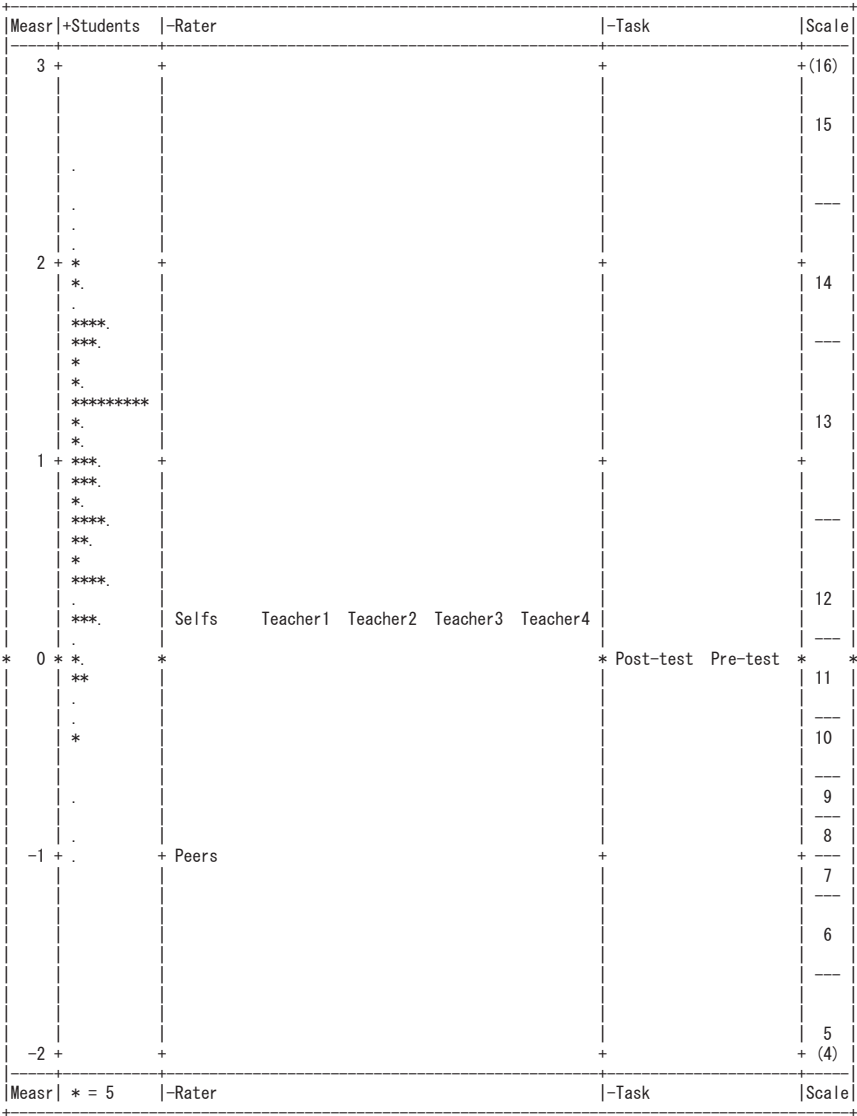
総合得点に関する信頼性

教員評価に比した生徒評価の信頼性に関する結果を提示する前に、MFRMの前提であるモデルの全体的適合度についてライティングの総合得点について計算をした。予想外の回答(Unexpected responses)における標準化残差を用い、 ± 2 を超えた標準化残差が約5%以内、 ± 3 を超えた標準化残差が約1%以内であれば、データがラッシュモデルに全体的に適合したと考えた (Linacre, 2021, p. 178)。総合得点については、データポイント数2,920でそれぞれが3.4%、1.4%であったので、全体としてこの基準は満たされていた。

図1の変数マップ(Wright マップ)は、縦軸Measr はロジットの尺度、Students は本研究に参加をした生徒、Rating methodsのSelfは自己評価、Peers はピア評価、Teacher 1～4は4人の教師評価それぞれを指す。変数マップの上位に位置するほど、評価が厳しいことを示し、下位に位置するほど評価が甘いことを意味する。また、Occasion はPre-testとPost-test は実験前後の英作文テストを意味し、変数マップの上に位置するほどテストが難しいことを意味し、下方にいくほど困難度が低いことを意味する。Scaleは評価表の合計点を示している。この結果によると、実験前後に行われた英作文テストの難易度はほぼ同じであり、生徒を意味するアステリスクがロジットの尺度0のやや上に集中することから比較的易しいテストだったことがわかる。また、4人の教師評価と自己評価がほぼ同じ所に位置することから、自己評価グループの生徒の評価の厳しさは教師評価と同程度だったと推定される。一方、ピア評価のグループの生徒は最も下方に位置するので、教師評価と自己評価グループの生徒に比べ評価が甘かったと言える。

次に評価の一貫性についてだが、インフィット及びアウトフィット統計量が評価者のラッシュモデルに対する適合度を表すので、こちらを参考に考える。表6が示すように、4人の教師全員のInfitとOutfitは、0.60～0.61に分布している。一般的には、0.7～1.3が適合を示すフィット統計量の基準だが(Bond & Fox, 2007)、本研究のテストで重要な判断を行わないことにより、0.5～1.5の時に、評価がモデルの予測値に適合していることを示すとした(平井 他, 2018)ので、教師評価は4人が全員ともに評価の一致度は適合度を示していた。教師評価、自己評価、ピア評価をそれぞれ分析した所、自己評価グループの生徒とピア評価グループの生徒についてはインフィットとアウトフィットが1.4以上を示すので、両評価タイプ共に評価は予測不可能で不一致度が高いと考えられる。つまり、受験者能力の低い受験者に高い点数を与えたり、受験能力の高い受験者に低いスコアを与えている可能性がある。なお、表は、評価者の厳格度の順に並んでいて、一番厳しい評価者は自己評価のグループの生徒で、一番評価が甘いのは、ピア評価のグループの生徒ということになる。また、教師評価者の人数が生徒評価グループの参加者よりも少ないことがこの結果に影響を与えたとも考えられる。

図1.
ライティング・テスト総合得点の変数マップ



注. Pre-testとpost-testはpre-task とpost-taskを指す。

表6.
総合得点に関する評価者特性

Raters	Observed Average	Measure logit	Model SE	Infit M Sq	Z Std	Outfit M Sq	Z Std
Self	12.42	.21	.05	2.91	9.0	2.94	9.0
Teacher 3	12.32	.20	.03	.60	-6.4	.61	-6.5
Teacher 1	12.33	.20	.03	.60	-6.4	.61	-6.4
Teacher 2	12.33	.20	.03	.60	-6.4	.61	-6.4
Teacher 4	12.33	.20	.03	.60	-6.3	.61	-6.4
Peers	13.97	-1.01	.05	2.44	9.0	2.57	9.0
Mean	12.76	.00	.04	1.29	-1.3	1.33	-1.3
SD (pop.)	.61	.45	.01	.99	7.3	1.02	7.3
SD Sample	.66	.50	.01	1.08	8.0	1.11	8.0

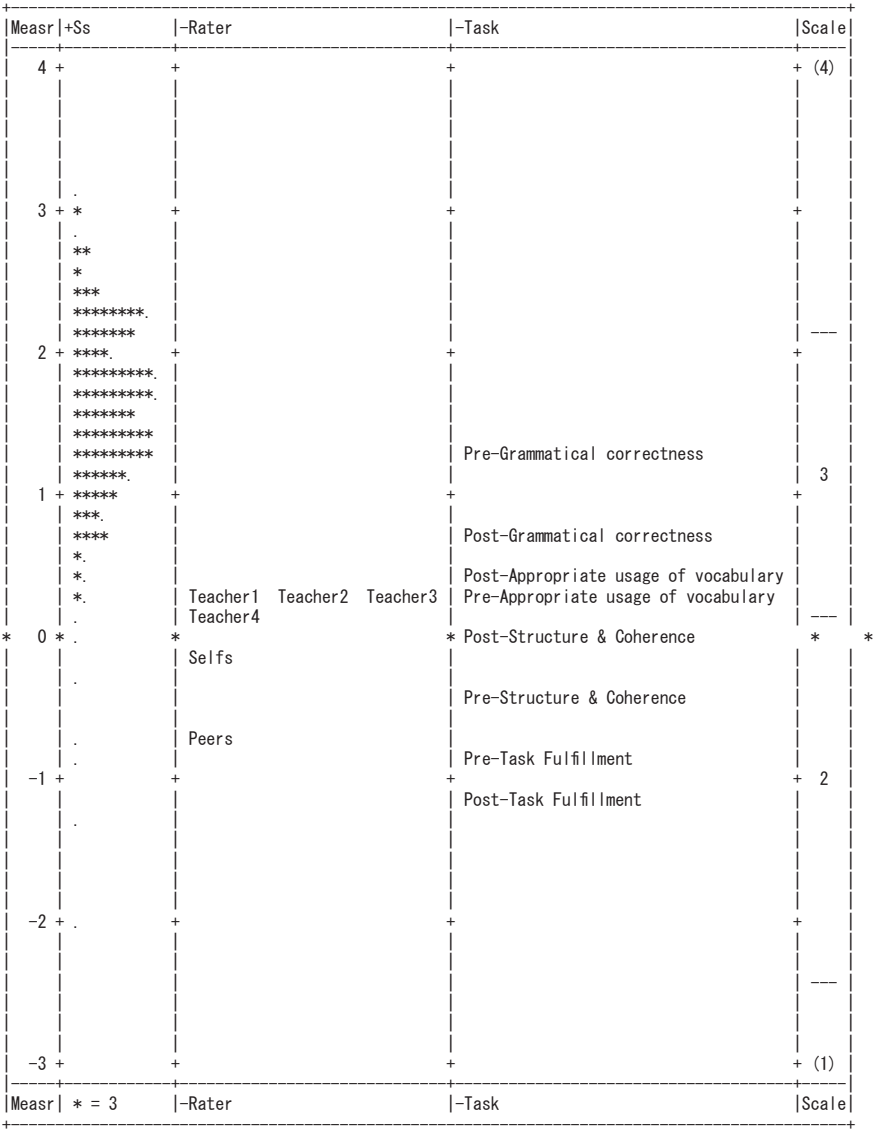
注. 下記はそれぞれの評価者タイプによるラッシュ分析の要約統計による。
全ての評価者(4人の教師、自己評価グループの生徒、ピア評価グループの生徒): Separation = 12.70; Strata = 17.26; Reliability = .99; χ^2 : 504.5; Significance = .00.
4人の教師: Separation = .00; Strata = .33; Reliability = .00; χ^2 : .00; Significance = 1.00.
自己評価グループの生徒: Separation = 1.85; Strata = 2.80; Reliability = .00 χ^2 : 1429.7; Significance = .00.
ピア評価グループの生徒: Separation = 2.06; Strata = 3.08; Reliability = .81; χ^2 : 679.3; Significance = .00.

分析的評価に関する信頼性

4つの項目からなる英作文の分析的評価の結果について、モデルの全体的適合度を調べたところ、予想外の回答(Unexpected responses) における±3を超えた標準化残差のデータポイント数が14、016の内、7.13%だった。よって、対数尤度カイ二乗検定(log-likelihood chi-square)を調べた所、対数尤度カイ二乗値が26、237.32(approximate model df = 13、709; p > .001)だったので、MFRMのモデルの全体適合度は満たされていると判断した。

図2は、実験前後の英作文テストの分析的評価の変数マップで、図の構図は、図1とほぼ同じであるが、四列目のOccasion * Rating scale (Occasion by Rating scale) は本研究実験前後の4つの観点に関する評価を表す点が異なっている。これは、Occasionによる実験前後それぞれの分析的評価のタスクの困難度を示す。第2列のテスト受験者である生徒の能力値は3.1から-2 logitを指しているが、大半は0より上なので難易度は低かったと考えられる。第3列の各観点の困難度の順番は、実験前の「正確な文法の使用」が一番高く、実験後の「課題達成度」が一番低かった。また、実験後の「正確な文法の使用」は他3つの実験後の各観点よりも困難度は一番高かった。要するに、4つの評価項目の中で一番困難度が高いのが「正確な文法の使用」で「課題達成度」が一番易しい項目であり、これは実験前後でも変わらなかった。左から第3列目の

図2.
ライティング・テスト分析的評価の変数マップ



評価の厳しさによれば、最も評価が厳しかったのはTeacher 1、次にTeacher 3、そしてTeacher 4と続く。生徒評価については、自己評価グループの生徒の厳格度が教師評価のそれに続き、最も評価が甘いのはピア評価グループの生徒の評価だった。この結果表示においては、図1と評価者の厳格度において教師評価と自己評価グループにおいてずれが見られるが、分析的評価では、Occasion ではなく、Occasion * Rating scale としたことが影響を与えたと考えられる。

次に分析的評価の一貫性についてだが、表7が示すように、4人の教師の4観点からなる分析的評価平均はいずれも3.06 もしくは 3.07で、分離信頼性(reliability statistics)が .00 を示しているので、評価の一貫性があり、おしなべて同程度の評価の厳格さがあると推定できる。対照的に、自己評価グループは、高い分離信頼性である 0.75 を示し、ピア評価のグループも 0.82を示している。評価の厳格度の一貫性をみるには、低い信頼性が好ましいので両評価タイプともに分析的評価の厳格度の一貫性はかなり低いと推定できる。さらに、両評価タイプともに、評価のばらつき (variability of severity)が自己評価グループでは1.75であり、ピア評価のグループでは2.16だったので、評価の厳格度にもばらつきがあると考えられる。

表7.
英作文の分析的評価に関する評価者特性

Raters	Observed Average	Measure logit	Model SE	Infit M Sq	Z Std	Outfit M Sq	Z Std	Separation ratio (G)	Separation (strata) index (H)	Separation reliability (R)
Teacher 2	3.06	.25	.03	.87	-5.2	.86	-5.1			
Teacher 3	3.06	.23	.03	.86	-5.5	.85	-5.4			
Teacher 1	3.07	.22	.03	.85	-7.4	.83	-7.7	.00	.33	.00
Teacher 4	3.07	.21	.03	.84	-6.4	.83	-6.3			
Self	3.23	-.15	.04	1.50	9.0	1.48	9.0	1.75	2.67	.75
Peers	3.45	-.77	.04	1.57	9.0	1.62	9.0	2.16	3.21	.82
Mean	3.21	.00	.03	1.08	-1.1	1.08	-1.1			
SD (population)	.17	.37	.01	.32	7.2	.34	7.2			

注. 下記はそれぞれの評価者タイプによるラッシュ分析の要約統計による。
全ての評価者(4人の教師、自己評価グループの生徒、ピア評価グループの生徒): Separation = 12.19; Strata = 16.58; Reliability = .99; χ^2 : 541.1; Significance = .00
4人の教師: Separation = .00; Strata = .33; Reliability = .00; χ^2 2 :1.7; Significance = .65
自己評価グループの生徒: Separation = 1.76; Strata = 2.67; Reliability = .76 χ^2 2 :1748.9; Significance = .00

ピア評価グループの生徒: Separation = 2.17; Strata = 3.22; Reliability = .82; χ^2 : 1764.2; Significance = .00

次に、評価者間の得点の一致率だが、教師評価が97.6% (17093/17520)を示す一方、自己評価グループは76.7% (20256/26421)、ピア評価グループは79.0% (17714/24966)だった。また、教師評価のフィット統計値は、0.84-0.87の範囲におさまるので、予測可能で一致度が高いと考えられるが、両生徒評価タイプともフィット統計値は1.5よりも高かった。

2. 自己評価とピア評価はどのようにライティング能力に影響を及ぼすのか。

英作文の総合得点に与える影響

研究課題2は、研究課題1で使ったラッシュアナリシスのデータを使う形で二つの評価法のライティング能力向上に対するそれぞれの影響について分析をしていく。表8は、実験前後の両評価タイプの総合得点の記述統計を表す。自己評価グループの生徒の平均点はピア評価のグループの生徒の平均点よりも実験前後ともにやや高いが近似している。標準偏差については、実験前後双方において、ピア評価のグループの方が自己評価グループよりも高い値を示している。

表8.

自己評価グループとピア評価グループの複合得点の記述統計

		人数	最低点	最高点	平均	標準偏差
事前英作文テスト	自己評価	147	4	16	12.27	1.64
	ピア評価	146	4	16	12.16	1.87
	合計	293	4	16	12.21	1.76
事後英作文テスト	自己評価	147	6	16	12.52	1.44
	ピア評価	146	6	16	12.33	1.62
	合計	293	6	16	12.43	1.54

しかし、表9の総合得点の課題測定レポート(task measurement report)によれば、自己評価の事前テストのmeasure logitは .33 であるが、ピア評価のグループの事前テストのmeasure logit は -.33 なので、事前テストは自己評価グループにとっての方がより難しかったといえる。事後テストについても同様の結果で、自己評価グループのmeasure logit は .05 であるが、ピア評価のグループのmeasure logit は -.05 だった。また、生徒全員のmeasure logit は事前テストが .12 で事後テストは -.12なので、両タイプともに生徒全体のライティング能力の向上が見られたと言えるが、自己評価グループの方がピア評価のグループよりもライティング能力が向上したと考えられる。

表9.
総合得点の課題測定レポート

Statistic	SA		PA		All students	
	Pretest	Post-test	Pretest	Post-test	Pretest	Post-test
Observed Average	12.27	12.17	12.71	12.25	12.23	12.43
Measure logit	.33	.05	-.33	-.05	.12	-.12
Model SE	.05	.04	.06	.04	.03	.03
Infit M Sq	.95	1.03	1.04	.93	1.00	.97
Z Std	-.8	.5	.6	-1.2	.0	-.7
Outfit M Sq	.92	1.03	1.04	.93	.99	.96
Z Std	-1.3	.5	.6	-1.3	-.1	-.8
Estm. Discrm	1.06	1.00	.95	1.02	1.02	.99

注. SA＝自己評価グループの生徒; PA＝ピア評価グループの生徒

英作文の分析的評価の得点に与える影響

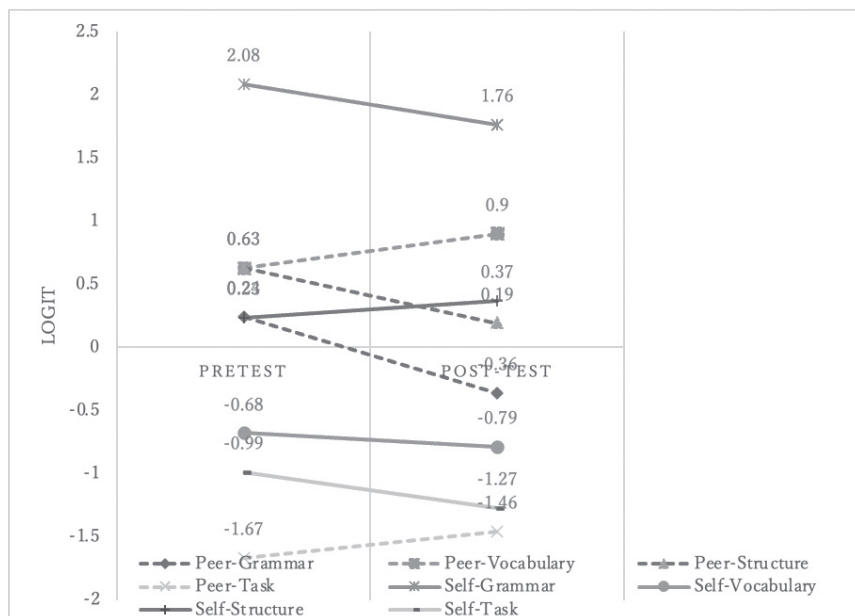
次に4つの観点における自己評価とピア評価の影響について考える。これについてもラッシュアナリシスのモデルの全体的適合度を調べたところ、一次元性 (unidimensionality) がラッシュ測定によって説明された変数の20%以上を占めたので十分に適合していた (Engelhard, 2013)。モデルの全体的適合度もデータポイント 9,344の内、11.5%が標準化残差の絶対値が3以上に関連していたので、対数尤度カイ二乗検定 (log-likelihood chi-square) を調べた所、対数尤度カイ二乗検定 (log-likelihood chi-square) の16,898 (approximate model df = 16,864.73; $p > .001$) が統計的に有意ではなかったのでラッシュアナリシスに適合すると考えた。課題測定レポート (task measurement report) に基づいて、事前事後テストの自己評価グループとピア評価グループの違いを分析するために図3を作成した (Appendix G参照)。

自己評価とピア評価グループ双方において課題の困難度から見た4つの観点は、事前事後テスト間で「構成と内容の一貫性」の観点以外の順位は変わらなかったが、事後テストは、事前テストに比べ、課題の困難度はやや変化をした。最も困難度が高かった観点は自己評価グループの「正確な文法の使用」だった。一方、最も困難度が低かった観点は、ピア評価のグループの「課題達成度」だった。図が示すように、自己評価とピア評価グループ両方において、「正確な文法の使用」のロジットが低下しているので、実験前に比べて「正確な文法の使用」の課題が両グループにとって簡単になったことを意味する。他の観点は、両グループとも事前・事後テストでロジットの増減が見られなかった。

表10は、自己評価とピア評価グループに関する課題測定の measure logits と Wald Statistics の結果を表す。この結果をもとに (1) タスク測定の measure logits; (2) Wald Statistics の二つの観点からライティング能力の向上についてそれぞれの評価タイプ

図3.

自己・ピア評価グループの4観点からなる分析的評価の事前事後テストの推移



注. Pre-testとpost-testはpre-taskとpost-taskを指す。

のライティングの能力の変化についての分析をする。なお、タスク測定に関する報告は、付録Gを参照。

まず、measure logits の結果についてだが、表10によると、自己評価グループのmeasure logits は事前テストの「正確な文法の使用」では2.08であり、事後テストでは1.76だった。同様に、ピア評価グループのmeasure logits は事前テストの「正確な文法の使用」では0.24だが、事後テストでは0.32だった。自己評価グループの事前事後テストの差は0.32で、ピア評価のグループの事前事後テストの差は0.60だったので、ピア評価のグループの方が差が大きかったと言える。

「課題達成度」のmeasure logits は自己評価とピア評価グループで違いを見せた。自己評価グループのlogit estimate は-0.99が事前の「課題達成度」の値であるが、事後テストの「課題達成度」は-1.27だった。一方、ピア評価のグループは、事前「課題達成度」は-1.67で、事後のそれは-1.46だった。端的に言えば、「文法の正確さ」と「課題達成度」において両グループ共に生徒評価の影響を受けたと考えられる。

次にWald Statistics を行い、それぞれの評価タイプにおける事前・事後テストの4つの観点それぞれの評価について統計的な有意差があるかを調べた。Wald Statistics は、有意水準0.05で臨界値1.96の時の帰無仮説を検定するものである。表10によれ

ば、自己評価グループでは、事前事後テストにおいて、総合得点と「正確な文法の使用」の観点で統計上の有意差があった。総合得点については、課題困難度の観点から見ると、1.44logits の差があった。Wald Statistics はこれを統計上有意だと確認した:t pre-composite, post-composite (283) = 5.63, $p < .05$, $d = .50$ 。よって、帰無仮説は棄却された。また、事前事後の自己評価グループの「正確な文法の使用」もまた帰無仮説は棄却された:t pre-grammar, post-grammar (113) = -2.74, $p < .05$, $d = .60$ 。つまり、自己評価グループの生徒は自己評価活動の後に総合得点と「正確な文法の使用」において課題困難度に違いを見せたということで、事前テストよりも課題困難度は低くなったので、能力の向上が見られたという意味になる。一方、ピア評価の生徒は、事前事後テスト間の「文法の正確さ」の観点において帰無仮説は棄却された:t pre-grammar, post-grammar (181) = -5.25, $p < .05$, $d = .40$ 。だが、自己評価グループの生徒とは異なり、総合得点の事前事後テスト間の帰無仮説は棄却されなかった。まとめると、自己評価グループは、ライティングの後に自己評価活動を行うとライティングの総合得点が向上し「正確な文法の使用」の観点も向上を見せた。一方、ピア評価グループの生徒は、総合得点の向上はできなかったが、「正確な文法の使用」においては向上をすることができた。

表10.
自己評価グループとピア評価グループの生徒に関するMeasure LogitsとWald Statistics

	Self-assessment group					Peer assessment group				
	Observed Average		Logit		Wald (pretest & post-test)	Observed Average		Logit		Wald (pretest & post-test)
	Pretest	Post-test	Pretest	Post-test		Pretest	Post-test	Pretest	Post-test	
Composite	12.27	12.71	.33	-.33	-5.63*	12.17	12.25	.05	-.05	-1.02
Grammar accuracy	2.33	2.64	2.08	1.76	-2.74*	2.22	2.74	.24	-.36	-5.25*
Vocabulary	2.98	3.07	-.68	-.79	-1.06	2.80	2.96	.63	.9	-1.61
Structure & Coherence	3.34	3.42	.23	.37	-.94	2.96	3.39	.63	.19	-.43
Task fulfilment	3.58	3.62	-.99	-1.27	-.47	3.52	3.68	-1.67	-1.46	-1.61

注. * $p < .05$

質的研究結果

生徒12名と教師二人を対象とした半構造インタビューの分析をもとに課題1と2についての分析をする。

自己評価グループ生徒のセオレティカル・コード

自己評価グループの半構造インタビューのデータは、第一段階で52のオープン・コードにまとめられ、それらは帰納的に16のアキシャル・コードに抽出された。これらは、10のカテゴリーに関連付けられ、それが4つのセオレティカル・コード(emergent themes)に分類された。この分類をもとに、カテゴリーとコード間の関係を理論化した。4つのセオレティカル・コードは、(A)評価基準と評価内容に対する意識化、(B)外部の存在、(C)英作文に対する学習態度、(D)英作文と自己評価によって促される情意、である。

(A)評価基準と評価内容に対する意識化

自己評価グループの生徒に対するインタビューの中で最も頻繁に登場したのが評価基準に対する意識づけに関する言葉だった。つまり、自己評価グループの生徒は、語彙、文法、英文構成や英文内容の一貫性に対する意識が高まったと英語力如何にかかわらず、強調をしていた。特に、語彙や文法を正確に使うように心がけたというコメントが多かった。こうした意識の高まりが英作文を校正し、書き終わった後に誤りをチェックする習慣の定着につながったようだった。例えば、英作文の得点が高かった生徒(SHSS1)は次のように答えている：

自分は英作文を書き終わる前に正確さをチェックする習慣が身に着きました。英作文を提出した後でさえ、自分が書いた英文の単語や文法の正確さを調べるために辞書を使って調べたり、友達に尋ねたりしました。(SHSS1)

このように、自己評価は評価基準に気をつけることによって校正をして見直す習慣づけにつながったと言える。また、英文構成や内容の一貫性についても自己評価グループ6人の内3人が英作文を書き終わった後に、書いた内容に筋が通っているか構成が整っているかを見直すようになったと答えている。英作文が高得点だった生徒(SHSS2)は次のように述べている：

自己評価をすることによって、私は英文と英文のつながりが自然かどうかを考えるようになりました。授業で習った英文構成に自分の英作文構成がしがたっているかを考えるようになりました。例えば、理由の述べ方や英作文での結論文の書き方などです。自己評価表は、自分の英文がうまく構成されているかどうか振り返るのに役立ちました。でも、自分一人で振り返るのは難しかったです。(SHSS2)

このように、SHSS2は、評価基準が英文校正や一貫性を見直す助けになったと認め、授業で学習したディスコース・マーカーを使って自分の英文を構成しようとした。以上のように、自己評価グループの生徒は、自己評価の時に使用した評価基準を語彙、文法、構成や一貫性を見直す道具として使い、その振り返りが英文を構成する意識の向上につながったと考えられる。二人の教師であるJETBとNETCも生徒たちが特に語彙や文法等の言語使用に注意を払い、それを見直すようになったと発言をしていた。また、JETBによれば、評価基準表が生徒たちに刺激を与え、生徒たちが現在知っている知識を最大限発揮しようという意欲につながったという。まとめると、自己評価グループの生徒は、自己評価の活動が評価基準の意識付け、特に語彙や文

法などの言語使用の正確さに対する意識の向上に良い影響を与えたと考えていることが分かった。

(B)外部の存在

自己評価グループでインタビューに答えた生徒の内半分が外部の存在である「読み手の存在」と「教師の評価」についての言及をした。英作文の点数がほぼ平均点だった自己評価グループの生徒(SMSS1)と英作文の点数が低かった生徒(SLSS2)は、自分が書いた英作文が読み手にとって理解可能かどうかについてわからないという不安を語った。自己評価をすることによって、生徒たちは自分の英作文に客観性があるかどうかや読者にとって理解が可能かどうかについて考えるようになった。今回の研究では、教師や友人が英作文を読むということは設定されていなかったからこそ、自己評価グループの生徒にとって外部の存在が浮かび上がってきたのかもしれない。

さらに、自己評価グループの生徒たちが外部の存在として、教師の存在の重要性や教師の指導の必要性を指摘した。これは、生徒たちが通常は教師の指導や助けを期待して英文を書いているので、それが無い時の不安感が吐露されたのである。また、自己評価の信頼性について疑いの気持ちをもつ生徒もいて、教師評価がこうした不安を解消してくれると英作文が平均点くらいだった生徒(SMSS2)は次のように訴えていた。

英作文の評価は先生によってされるべきだと思います。なぜならば、私は自分の英語力に自信がないからです。自分の英語が正しいかどうか判断するのは自分にはとても難しかったです。(SMSS2)

また、インタビューに答えた半分の生徒が自己評価の効果を認めながらも、読者がいないので自分の考えを共有できない不満や自己評価が自己満足で終わってしまう不安もコメントしていた。英作文の点数が低かった生徒(SLSS2)は次のように述べている。

私は英作文は読む人にとってわかりやすくなければいけないと思います。でも、自分の英語が他の人にとって意味が通じるものなのか自分ひとりでは、よくわかりません。ひとりでは、自分の英語が理解可能かどうかは想像できません。(SLSS2)

本研究では、自己評価の活動に教師や友人のフィードバックを介在させなかったことで、教師などの外部の存在に対する意識が生徒たちの中でより強くなったようである。特に教師は生徒たちに信頼のおけるフィードバックや新しい知識を与える存在としてその必要性が強調されていた。インタビューを受けた二人の教師も客観性や学習意欲が向上をするので、読者の存在が生徒たちに与える効果や重要性について言及をしていた。

(C)英作文に対する学習態度

英作文に対する学習者の意識は、「英文を書くことに対する態度」や「学習者の考

え方に対する効果」の観点から浮き彫りになった。インタビューに答えた6人全ての自己評価グループの生徒が自己評価は自分の英文を書くことに対する態度に良い影響を与えたとコメントをした。例えば、英作文の点数が高かった生徒(SHSS1)は、現在の自分の能力と学習目標の間にあるギャップを理解させてくれたので、自己評価が自分の英語を書く学習態度が改善されたと次のように言及をしている。

自己評価をすることによって、自分の弱点がわかったのもっと英語を勉強しようとする気が出ました。今のレベルと最終ゴールのギャップを埋めなければだめだと思いました。自己評価をすると、自分を振り返ることができます。それがもっと勉強しようという気持ちにさせるのです。(SHSS1)

一方で、6人の内2人の生徒が学習の方向性を見失ったと報告をしている。例えば、英作文の点数が平均程度だった生徒(SMSS2)は教師の指導がなかったことや他者からの導きがなかったことによって不安を感じたと述べている。

自分が何をすべきなのかわからなくなってしまったので自己評価について不安を感じました。誰かの助けがあれば、もっと自信をもって英作文を書いて自己評価をすることもできたと思います。(SMSS2)

このように、自己評価の英作文に向かう態度に良い影響をインタビューに答えた全員が認めながらも一部の生徒は英語を書くことに対する意識についての負の影響も指摘していた。インタビューに答えた教師も自己評価が生徒たちの英作文に対する意識に正と負の影響の両方を与えると述べていた。JETBは、正の影響としては、自己内省を深めることによって、自己校正力を向上させ、より適切に英文を書いていけるようになる」と指摘している。また、NETCは、ほとんどの生徒たちが自分の考えや経験を英語で書くことを楽しみ、英語で書くことを価値ある体験だと認識していると述べている。一方、学習者の意識に及ぼす負の影響としては、JETBとNETCの両者ともに生徒によっては自己評価が学習に向かう姿勢を消極的にさせる原因にもなったとコメントした。本研究は、10日間の短期間に集中的に教師や友達の助けや指導がない形式の実験だったので、こうした負の部分が解決策のないまま、一部の生徒の情意の中で増強されたのかもしれない。よって、自己評価が生徒に与える負の影響を防止するためには、こうした生徒のストレスを緩和することが必要だ。そのためにも教師の助けや指導が良いタイミングで必要とされると二人の教師は回答をしていた。まとめると、自己評価は生徒の意識や学習態度に概して良い影響を与えるが、欲求不満などの負の影響を与えることもあるということがわかった。また、それらを抑止するためにも教師の適切なガイダンスが期待されている。

(D)英作文と自己評価によって促される情意

インタビューに答えた6人の内4人の自己評価グループの生徒が自己評価をすることが自分に対する自信につながったと答えた。具体的には、課題を完成したときに感じる自己達成感や十二分に努力をして課題をやり遂げたことに対する充足感を指摘した生徒が多かった。英作文が平均点程度だった生徒(SMSS1)は、英作文は苦手だったが今回の活動で努力をしたことに触れ、達成感を感じたとコメントをしている。

英語は得意ではないので、英語で自分の考えを書くということはとても難しかったです。でも、自分は課題に対して頑張って取り組みました。ただ、自己評価に使ったルーブリックには、どれだけ努力したかを書き込む欄がなかったので、少し残念でした。でも、とにかく頑張って、努力をして英作文を書きました。だから、今は英作文にたいして苦手意識はありません。(SMSS1)

また、インタビューに答えた自己評価グループの6人の生徒の内4人が英語の勉強に対する意欲が高まったとコメントをしている。英作文で高い点数であった生徒(SHSS2)は自己評価活動のおかげで英語の勉強への動機付けとなり、学習方法を再考するようになったと答えている。しかし、英作文の点数が低かった生徒(SLSS2)は、自己評価活動は、自分の現在の能力と目標のレベルのギャップを埋めることはできず、かえって自己評価活動が彼女をいらだたせることになったという。つまり、自己評価活動が、自分の英語力の欠如を際立たせることになり、英作文を書くことに対して意欲が減退したとコメントしていた。よって、この生徒は自己評価活動について負の感情を抱いていると述べている。

私も英語で文章を書こうと努力をしたし、自分の英作文を見直し評価をしようと思いました。でも、自分の英語のできなさは全く変わらなかった。とにかく、(自己評価をしても)英語は不得意なままであることがよくわかりました。だから、英作文に対して良い気持ちは持っていません。(SLSS2)

インタビューに答えた二人の教師も、自己評価活動と英作文を集中的に書いたことが生徒の情意に与える正と負の影響について報告をしていた。NETCによれば、自己評価は生徒によってその情意に与える影響は異なり、英語力が高い生徒は肯定的に自己評価の結果を受け入れるが、英語力が低い生徒は劣等感を強く感じる傾向があるとのことだった。JETBもこのNETCのこの観察に同意をしていて、自己評価が自分の成果と目標値との距離を自覚させるので、英語力が低い生徒は自分の置かれた厳しい現実にはやる気が萎えてしまうのかもしれないと答えていた。以上のように、自己評価は英語ライティングに対して自己達成感や満足感、そして意欲を高めることもできるが、生徒によっては、学習目標と自分のレベルの開きに目覚めストレスを感じることもあることがわかった。

ピア評価グループ生徒のセオレティカル・コード

ピア評価のグループの生徒のインタビューからは、最初のコーディングから38のオープン・コードを抽出し、それを22のアクシシャル・コードに分類することができた。また、これらは8つのセオレティカル・コードにまとめられた。8つのセオレティカル・コードの関係を分析するために、それぞれのカテゴリーの核となる要素と概念を明確に表現している部分をコード化し、続けて互いの関係に注意を払いながら、3つのセオレティカル・コードを決定した。8つのセオレティカル・コードは、以下の3つのセオレティカル・コードに抽出された：(E) 友達からの学習効果、(F) メタ認知の発達、そして(G) 英語学習に対する意識。

(E) 友達からの学習効果

ピア評価のグループの生徒へのインタビューから最も頻繁に言及されたのは友達から受ける影響についてだった。インタビューを受けた6人全員の生徒が友達の英作文から何かしらを学ぶことができたコメントをしていた。6人の内、5人の生徒が友達の英作文を言語使用、構成、課題達成、そして英文内容の独創性の観点から賛辞を送っていた。さらに、6人全員が友達の英作文を読むのは楽しく、英作文を読むことによって友人関係が深まり、友達の考えや経験を知ることができるのは喜びであると語っていた。また、友達の英作文を評価するのも楽しいと述べていた。これが、ピア評価のグループの生徒のインタビューの最も際立つ特徴だった。よって、生徒たちは、ピア評価活動を通して、友達とのインタラクションを確立することができたことを肯定的に受け止めていた。例えば、英作文の評価が低かった生徒(PLSS1)は、友達の英作文を読むことによって、友達の考えを知ることができたとして次のように述べている。

友達の考えや興味を発見できるので、クラスの人たちの英作文を読むのはとても楽しかったです。同じクラスでもあまり話す機会がなかった人たちと会話をするチャンスになりました。英作文の授業やピア評価活動が終わった後も、もっと話がしたいなと思いました。(PLSS1)

また、ピア評価のグループの生徒6人の内4人の生徒が異なる意見や考え方を知ることができるので、他者の見方を共有する大切さについてコメントをしていた。英作文の評価が平均点レベルである生徒(PMSS1)は、ピア・リーディングによって友達から英語の書き方を学ぶことができるし、新しい知識を得ることもできると言及している。

友達の書いた英文を読んだおかげで、新しい単語や使い方を知ることができて良かったです。英文をどう構成するかを学ぶこともできました。だから、次に英作文を書くときには、友達の英文の書き方をまねたいと思っています。友達が頑張っているの、自分も頑張らなきゃという気持ちにとてもなりました。(PMSS1)

PMSS1の言葉は、ピア評価が生徒に、友達、読者、評価者、そして書き手という複数の役割をする機会を与えるということを示唆している。言い換えれば、ピア評価をした生徒は、多くの視点で友達の英作文を読んで評価をしているのである。こうした体験が英文の書き手としての成長に肯定的な影響を与えていると考えられる。この点については、二人の教師も同意をしていた。特に、JETBは、ピア評価を授業の学習ツールとして有効であるとコメントをしている。つまり、ピア評価はライティングの授業を活発な雰囲気になせ、生徒たちを協同的に学ばせることを可能にさせるのだ。

(F) メタ認知の発達

「メタ認知の発達」というセオレティカル・コードは、「読者の存在を意識化すること」と「読み手にとって読みやすい英文を書くことを意識化すること」の二つのコードから構成される。なお、メタ認知とは、自分の認知活動を客観的にとらえることを意味する(三宮、2018)。まず、「読者の存在を意識化すること」については、ピア評価グループのインタビューに答えた6人の生徒全員が、自分の英作文を読む友達の存在を意識

して書くようになったと答えていた。英作文の評価が平均点レベルだった生徒(PMSS1)は、友達を読むことを強く意識して、例示を工夫するなどして説得力がある英文になるように努力をしたとコメントをしていた。

クラスメイトが自分の英文を読むことがわかっていたので、わかりやすい英文を書くために注意を払って書こうと思いました。特に、友達に自分の考えや体験をわかってもらうために客観的に書こうと思ったので、具体的な例を出そうと思いました。ピア評価をする時には、友達の英文を読みながらも自分の英作文のことを思い出して、頭の中で(自分の英作文と友達の英作文を)比較をしました。(PMSS1)

また、英作文の評価が高い生徒 (PHSS2)は、友達から自分の英作文をほめてもらい学習意欲が高まったと述べている。さらに、この生徒は、友達が評価をするので、英作文を書く時の意識に変化があったとも言及した。この意識の変化こそが、他者の存在の認識であり、それを意識するメタ認知の発達につながったと考えられる。

ピア評価はいつも自分に読む人のことを考えさせてくれました。ピア評価がなければ、自分が書いた内容は違うものになっていたと思います。友達が自分の英文を読んでくれることやピア評価が良いことかどうかはわからないが、書いている間、クラスメイトの誰かがこれを読むということはいつも意識していました。自分の英作文を読んで良い評価をくれるといいなと思っていたし、自分が評価を意識していることも自覚していました。(PHSS2)

これについてJETB は、ピア評価活動の最中に生徒が、書き手や読み手、評価者、そして友達などの立場を自然に変化させていることがメタ認知の発達を導いたとコメントをしている。つまり、様々な役割を担っていることが、生徒のモニターの意識を目覚めさせ、ライティングの力を向上させると語っている。このように、ピア評価は、学習者にライティングや評価活動をモニタリングさせることが可能で、それがメタ認知の発達に寄与すると考えられる。

(G)英語学習への意識変化

ピア評価に参加をしてインタビューを受けた6人の生徒全員が、ピア評価が英語で書くことに与える心理的影響について語っていた。また、肯定的影響を意味するコード数の方が、否定的な影響のコード数よりも多かった。肯定的な言及としては、ピア評価が英語学習に対する意欲を高め、意欲的に英作文を書きたいという刺激を与えたということだった(PMSS2;PHSS1)。この刺激は、ライティングのみならず、リーディングや英語の各種検定試験、そして大学入試に対する意欲にもつながったとコメントをしている(PHSS1)。また、友達の英文を読みフィードバックをもらうことによってクラスに良い雰囲気醸造されることを高く評価している生徒もいた(PMSS1)。

一方で、ピア評価が生徒の情意に与える負の影響についてコメントをしている生徒もいた。例えば、ピア評価の難しさについて、英作文の評価が平均点のレベルだった生徒(PMSS2)は以下のように述べている。

ピア評価は、いつもではないが、時々、勉強への意欲を刺激してくれるし他の人からたくさん学ぶことができるので良いと思います。でも、課題(英作文)を全く完成していない人もいますので、そういう人の英作文を読み評価をする場合は、どう評価をしてよいのかわからなくて困りました。(PMSS2)

このようにピア評価の意義について疑問を投げかけている。また、ピア評価は評価をする友人との人間関係に影響をされるので、良い評価を友達からもらってうれしいが、果たしてその評価は正直なものなのか疑問を抱くともPMSS2はコメントしている。別の英作文評価が低い生徒(PLSS2)は、ピア評価の英語ライティング向上に効果的かどうかについても疑問を呈した。なぜならば、彼女自身が自分の英語力に不安を感じていて友達の英語を読み信頼性のある評価ができる自信がないとコメントをしているからだ。

私は英語が不得意なので、他の人の英作文を判断する能力はないと思います。だから、自分のピア評価は良いものではないし正しくないと思います。だから、評価をされた人には悪いなと思っているので、ピア評価は好きではありません。自分には、他の人の英文が正しいのかわからないし、一人では判断するのは難しかったです。(PLSS2)

このように、ピア評価は個々の生徒の英語力や英語に対する態度、そして生徒同士の関係性によって影響を受けるようである。これらの生徒によるインタビュー内容に呼応するように、二人の教師もピア評価には正の局面と負の局面があるので、授業への導入には注意が必要だと語っていた。JETBは、ピア評価を授業で扱うときには、特にクラスの間人間関係に気配りをし、ピア評価の前に簡単なルールを設けてやる気をなくす生徒が出ないようにしていると述べている。

ピア評価は生徒に競争意識を持たせ、生徒の客観性ややる気を向上させることができます。でも、クラス内の人間関係に対する配慮は必要でしょう。匿名でのピア評価をするのも一つの方法でしょう。ピア評価をする前に、生徒には、他者に意見を言うときに注意をしなければならないこと、悪い点ばかりではなく、必ず良い点を見つけてあげてそれを言うようにと指導をしています。(JETB)

このように、教師達もピア評価をより実りあるものにするように努力をしていることがわかる。NETCも生徒は教師からの言葉や評価よりも友達の評価の方を深刻に受け止める傾向があると言及している。よって、教師たちは、ピア評価が時に生徒同士の関係に影響を与えることも考えて注意を払うべきだろう。以上のように、ピア評価がライティングへの意識に与える影響は、教師による配慮や導入前のルールで肯定的なものになるようだ。

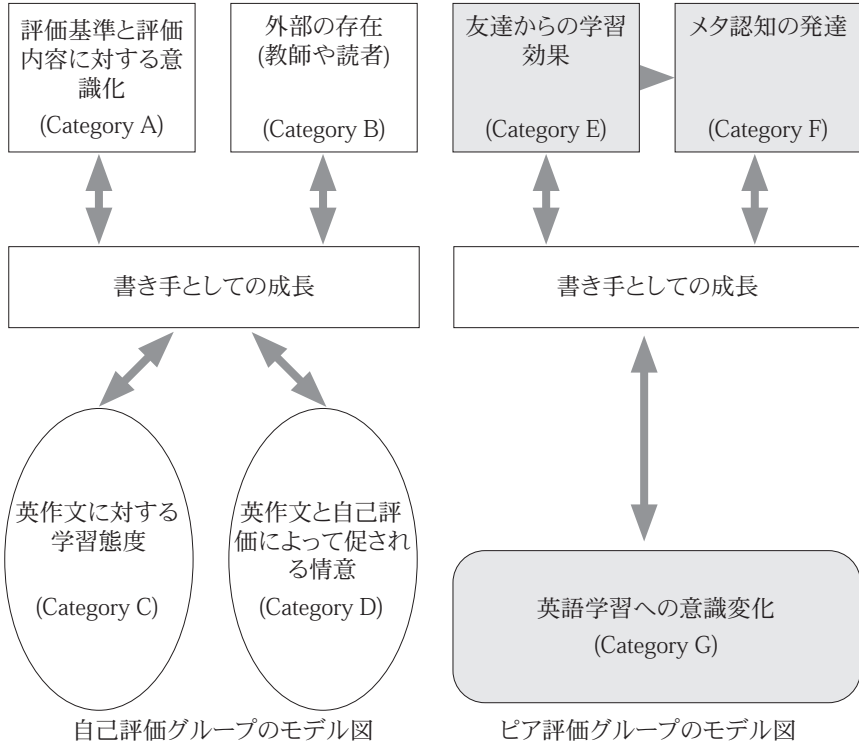
半構造インタビューの結果から見た自己評価とピア評価の比較

Corbin と Strauss (2015)によると、コーディングの過程でコードの互いの関係性に着目してダイアグラムを作成することがコーディングの関係性や秘匿された重要なテ

ーマを暴くので、図4のように自己評価とピア評価がライティング活動に与える影響を二つのモデルにまとめた。

図4.

自己評価グループとピア評価グループモデル比較



自己評価とピア評価の生徒に共通するのは、学習者評価を通じて両タイプの生徒が英語の書き手としての成長を目標(コア)に置いていることである。さらに、サブ・カテゴリー(セオレティカル・コード)の矢印は全てこの目標(英語の書き手としての成長)を目指している。一方、自己評価とピア評価のモデルで違う点は、サブ・カテゴリーの「外部の存在」の意味が異なっている。これは、自己評価では、外部の存在は物理的にはないが、ピア評価では友人からの影響を直接受けることを意味している。他の相違点としては、ピア評価モデルにはメタ認知の発達が提示されているが、一方、自己評価モデルには自己評価活動が与える自己達成感などを意味する情意が提示されている。

考察

ここでは、量的研究結果と質的研究結果を統合し、まず二つの研究課題を考察し、その後に研究課題3である「学習を促す評価(LOA)」における二つの学習者評価の役割を考える。最初の研究課題である学習者評価の信頼性についてであるが、量的研究結果の通り、厳格さにおいては、自己評価は教員評価と同等の厳しさを示した。これは、質的研究結果も、自己評価活動が英作文の評価基準や評価項目に対する理解を深めるのに有効であったことを示すので、自己評価は評価の厳格さにおいては信頼性の高い評価法といえよう。だが、量的研究結果によれば、自己評価は一貫性には乏しく、質的研究結果の生徒からの発言からも、生徒たちは自己評価に対する自信が不足気味であることがわかったので、質的研究結果は量的研究結果を支持しているといえる。先行研究によれば、自己評価の信頼性は研究対象によって若干の差異はあるもののある程度高いとされているが、年齢が若い学習者の場合は評価の信頼性に揺れが生じると言われているので(Esfandiari & Myford, 2013)。本研究も若い学習者を対象としたので、信頼性に年齢が影響を与えたと考えられる。また、本研究は、信頼性を厳格さと一貫性の観点から分析したので、自己評価の信頼性の新しい側面がわかったといえる。

一方、ピア評価は、量的研究結果によれば、教師評価や自己評価にくらべ、評価は甘く、評価の一貫性も見られなかった。質的研究結果の半構造的面接からは、ピア評価は生徒の英語力や生徒同士の人間関係にも影響を受けていることが観察され、ピア評価の信頼性を疑問視する意見も一部の生徒から発声された。よって、量的と質的ともに本研究におけるピア評価の信頼性については低いと考えられる。ピア評価の信頼性は高めであるとする先行研究結果が多いが、その多くは大学生を対象とした研究である。日本の高校生を対象とした分析結果は少ないので、若年学習者によるピア評価の信頼性が新たにわかったといえる。

次に、第二の研究課題である生徒評価のライティング能力への影響に関してだが、自己評価とピア評価は共に文法の正確さに良い影響を与えるが、その影響の質が異なることが考察された。上記のように、厳格さにおいて教師評価と同等の自己評価は英作文の評価基準に対する認識を高め、自己統制や内省力を深めると考えられる。質的研究結果が示すように、自己評価によって醸成されたその強い自己認識は、語彙や文法知識の不足を意識することにおいて特に強く現れていた。しかし、こうした強い自己認識や内省は、英語力に対する自信喪失にもつながることもあるが、文法の正確さを向上させるために肯定的な影響を与えた。さらに、自己評価グループにおいては、総合得点の向上も見られた。また、質的研究結果の二つの評価法を比較したモデル(図4)が示すように、自己評価グループの生徒の方が、ピア評価グループの生徒よりもより評価基準や評価内容について頻繁に言及をしていたので、評価基準をより強く意識し、その気づきが英作文の全体的向上に役立ったと考えられる。生徒評価の学習効果としてライティングにおける校正能力への向上が先行研究でも報告されている。本研究の結果でも英文を書いた後に校正する習慣に結び付いたという結果が出ているので、先行研究の結果と同じである。

一方、ピア評価においては、生徒同士で英作文を読み合い評価をすることによって、同級生との交流を楽しみ、友達が書いた英語から学ぶことが文法力向上の助けになったと思われる。ピア評価のグループの生徒たちの多くは、友達の英作文を読む楽しさを指摘し、大半の生徒が肯定的にピア評価やピア・リーディングを受けとめてい

た。自己評価グループの生徒たちよりも自然に楽しんで評価活動を行っていたことが質的研究結果から観察された。友達同士で励ましや誉め言葉を交換することが、教室内の雰囲気や和やかにし、生徒間の交流を深め、そうした協同的学習スタイルが英作文に取り組む態度に効果的な影響を与えたと考えられる。また、ピア評価を通して、書き手、評価者、読み手、友達等の多様な役割を自然に担うことが、メタ認知の発達を促したことも、友達の英作文から学ぶことを効果的に促したと考えられる。

次に、第三の研究課題である「学習を促す評価(LOA)」における二つの学習者評価の役割について考察をする。先行研究が示すように、自己評価とピア評価の双方とも、普段は学ぶ立場の学習者が立場を変えて評価をする役割に代わることによって、学習者を学びの中心に置くことができるので(Leahy & Wiliam, 2012)、ライティング授業の活性化につながり、学習者による評価は「学習を促す評価(LOA)」の目的に合致すると考えられる。また、信頼性には欠けていたが、二つの評価方法ともに学習者に評価基準や評価内容に関するメタ認知力を上昇させた。それによって、学習者にとって何が必要で何が不足しているのかを自分で考えさせることができた。これが、学習者のメタ認知の発達や自律的学習の助けとなり、学習を促していくことになることが本研究では明らかになった。また、教師にとって適切なフィードバックをするための重要な情報となりうることも本研究の結果によってわかったことである。本研究では、教師は観察者で評価活動の介在やその後のフィードバックはしなかったが、質的研究結果によれば、生徒は学習者による評価だけではその信頼性に対し不安を感じ、教師によるフィードバックを期待していることがわかった。また、知識の不足を自覚した時に、次の学習段階に進むための助け(scaffolding)も望んでいた。自己評価によって内省し次に何をすべきかを自覚することはできる。また、友達から賛辞をもらっても、具体的に次にどう前に進めばライティング能力が向上するのかを具体的に助言してくれる友達は少ないことがピア評価グループの生徒は不満だったようだ。このように、フィードバックのみならず、前に進ませる(feed-forward)指導や助言こそが、生徒が教師に求めることだろう(Hamp-Lyons, 2017)。よって、「学習を促す評価(LOA)」をうまく機能させるためにも、学習者による評価によって、学習者を評価の中心に置きながらも、教師による適切な介在は必要だろう。これは、学習者による評価のみではその信頼性において不十分であり、たとえ学習者の自立度が向上したとしても、学習者の生徒評価に対する自信や信頼性には不安が残るためである。教師は、今後の学習に必要なことを示す役割が求められていて、自己評価とピア評価の結果は、具体的なその判断を下す資料となる。これまでの「学習者による評価」に対する先行研究は、生徒評価が教師評価の代替評価として可能かということが主要な研究課題の一つだったが、学習者による評価の本質は、学習を前に推し進めるための重要な情報を教師に提供することにあると考えられる。

結論

本研究は、日本の高校生を対象に、自己評価とピア評価という二つのタイプの評価法の信頼性やライティング能力への影響を探究することを目的とした。二つの評価方法を比較することによって、両者の類似点や相違点を明らかにし、それぞれの特質を具体化し「学習を促す評価(LOA)」を進める足掛かりとすることも目的とした。生徒評価は単なる教師評価の代替ではなく、教師評価とは異なる視点や特質をもつ評価法であるので、教師はその特質を理解し授業に採用することが求められている。本研究

は、自己評価とピア評価各々の特徴を明らかにするために、できるだけ他の教育的影響を避ける目的で、短期間に集中的に実験を行った。しかしながら、長期的な視点が不足するために、「学習を促す評価(LOA)」が目指す学習の永続性に関する分析と考察が不足している点が限界点だと言える。また、それぞれの評価タイプのデータ数に違いがあるので、それが誤差の大小に影響している可能性がある点も限界点だと言える。今後の研究としては、生徒評価へのフィードバックやフィードフォワードを教師がどのようにいづすべきかについての分析や生徒評価の信頼性と英語力の関係についての分析、さらに、自己評価とピア評価を両方取り入れたハイブリッド形式の効果や適切な運用方法についてさらに進めていく必要がある。

注. 付録は https://osf.io/nuvgx/?view_only=f1699ee01b454d029ee48e94b007edfeを参照。

All appendices are available from the online version of this article at <https://jalt-publications.org/jj>.

すべての付録は、<https://jalt-publications.org/jj>のオンライン版から入手できます。

謝辞

本調査を実施するにあたり、参加者の皆様には多大なご協力をいただきました。また、原稿執筆にあたっては、査読者の皆様から貴重なご助言をいただきました。心より感謝申し上げます。本研究は一部著者の博士論文に基づいています。

著者略歴

大井洋子は、神奈川県で高校教諭として英語教育に携わり、早稲田大学大学院にて博士号を取得。現在は清泉女子大学言語教育研究所で言語教育コーディネーターとして言語教育に携わっている。

参考文献

- 戈木クレイグヒル滋子(2018).『グランデッド・セオリー・アプローチ 改訂版 理論を生み出すまで』新曜社.
- 三宮真智子(2018).『メタ認知で＜学ぶ力＞を高める:認知心理学が解き明かす効果的学習法』北大路書房.
- 平井明代・横内裕一郎・加藤剛史(2018).「項目応答理論:標本依存と項目依存を克服した測定を実現する」平井明代編『教育・心理・言語系研究のためのデータ分析—研究の幅を広げる統計手法』(pp. 94-137)東京書籍.
- 行森 まさみ(2018)「学習指導要領における英語4技能再考」『表現学』第4号 (pp.10-17) 大正大学表現学部表現文化学科.

- Alsowat, H. H. (2022). An investigation of Saudi EFL teachers' perceptions of learning-oriented language assessment. *European Journal of English Language and Literature Studies*, 10(3), 16–32. <https://doi.org/10.37745/ejells.2013/vol10no3pp.16-32>
- Aslanoglu, A. E., Karakaya, I., & Sata, M. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet Rasch model. *Eurasian Journal of Educational Research*, 89, 25–46. <https://doi.org/10.14689/ejer.2020.89.2>
- Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education*, 17(2), 199–214. <https://doi.org/10.1080/09695941003696172>
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Open University Press.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, 86(1), 8–21. <https://doi.org/10.1177/003172170408600105>
- Bond, T. & Fox, C. (2015). *Applying the Rasch model fundamental measurement in the human sciences*. Routledge.
- Boud, D. (1992). The use of self-assessment schedules in negotiated learning. *Studies in Higher Education*, 17(2), 185–200. <https://doi.org/10.1080/03075079012331377621>
- Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, 18, 529–549. <https://doi.org/10.1007/BF00138746>
- Butler, Y. G. & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27(1), 5–31. <https://doi.org/10.1177/0265532209346370>
- Carless, D. (2007). Learning-oriented assessment: Conceptual basis and practical implications. *Innovations in Education and Teaching International*, 44(1), 57–66. <https://doi.org/10.1080/14703290601081332>
- Chang, W.-C., & Chan, C. (1995). Rasch analysis for outcomes measures: Some methodological considerations. *Arch Phys Med Rehabil*, 76, 934–939. [http://www.archives-pmr.org/article/S0003-9993\(95\)80070-0/pdf](http://www.archives-pmr.org/article/S0003-9993(95)80070-0/pdf)
- Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4), 891–901. <https://doi.org/10.1037/0022-0663.98.4.891>
- Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

- de Wever, B., van Keer, H., Schellens, T., & Valcke, M. (2011). Assessing collaboration in a wiki: The reliability of university students' peer assessment. *The Internet and Higher Education*, 14(4), 201–206. <https://doi.org/10.1016/j.iheduc.2011.07.003>
- Derakhshan, A., Shakki, F., & Sarani, M. A. (2020). The effect of dynamic and non-dynamic assessment on the comprehension of Iranian intermediate EFL learners' speech acts of apology and request. *Language Related Research*, 11(4), 605–637. <https://lrr.modares.ac.ir/article-14-40648-en.html>
- Dieten, A. J. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing*, 6(1), 30–46. <https://doi.org/10.1177/026553228900600105>
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.
- Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*, 18, 111–131. <https://doi.org/10.1016/j.asw.2012.12.002>
- Engelhard, G. (2018). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
- Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3), 287–322. <https://doi.org/10.3102/00346543070003287>
- Farrokh, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34(1), 79–102. <https://doi.org/10.37546/JALTJJ34.1-3>
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine. <https://doi.org/10.1097/00006199-196807000-00014>
- Green, A. (2017). Learning-oriented language test preparation materials: A contradiction in terms? *Language Testing and Assessment*, 6(1), 112–132. <https://doi.org/10.58379/SFUN3846>
- Hamp-Lyons, L. (2017). Language assessment literacy for learning-oriented language assessment. *Language Testing and Assessment*, 6(1), 88–111. <https://doi.org/10.58379/LIXLI198>
- Jones, N., & Saville, N. (2016). *Learning oriented assessment*. Cambridge University Press.
- Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 49–72). Sage. <https://doi.org/10.4135/9781446250808.n4>

- Lee, H. (2017). The effects of university English writing classes focusing on self and peer review on learner autonomy. *The Journal of Asia TEFL*, 14(3), 464–481. <https://doi.org/10.18823/asiatefl.2017.14.3.6.464>
- Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.
- Linacre, J. M. (2021). *A user's guide to FACETS Rasch-model computer programs: Program manual 3.83.5*. <http://www.winsteps.com/manuals.htm>
- Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26(1), 75–100. <https://doi.org/10.1177/0265532208097337>
- McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice*, 10(2), 209–220. <https://doi.org/10.1080/0969594032000121289>
- Ministry of Education, Culture, Sports, Science and Technology. (2011). *Wagakuni niokeru kokuren jizoku kano na kaihatsu no tameno kyoiku no ionen jisshi keikaku* [Education for sustainable development for our country]. <https://www.mext.go.jp/unesco/004/detail/1359220.htm>
- Ministry of Education, Culture, Sports, Science and Technology. (2018). *Heisei 29 nenndo eigokyoiku chosa kekka* [The results of fiscal 2017 survey of proficiency in English]. https://www.mext.go.jp/component/a_menu/education/detail/__ics-Files/afieldfile/2018/04/06/1403469_02.pdf
- Mok, M. M. C. (2012). Assessment reform in the Asia-Pacific region: The theory and practice of self-directed learning-oriented assessment. *Examinations Research*, 4(33), 79–89. https://doi.org/10.1007/978-94-007-4507-0_1
- Mulvey, B. (2016). Writing instruction: What is being taught in Japanese high schools, why, and why it matters. *The Language Teacher*, 40(3), 3–8. <https://doi.org/10.37546/JALTTLT40.3-1>
- Oi, S. Y. (2018). The relationship between writing tasks in textbooks and can-do lists in terms of task difficulty. *Journal of Pan-Pacific Association of Applied Linguistics*, 22(2), 53–70. <https://doi.org/10.25256/PAAL.22.2.3>
- Oi, Y. (2019). Japanese high school English teachers' perspectives on classroom writing assessment criteria: A needs analysis. *The Bulletin of the Graduate School of Education of Waseda University*, 27(1), 159–176. <http://hdl.handle.net/2065/00063295>

- Oi, Y. (2021). *Efficacy of student assessment as part of English writing instruction for Japanese high school student* (Publication No. 6413) [Doctoral dissertation, Waseda University]. National Diet Library. <https://iss.ndl.go.jp/books/R100000039-1003071522-00>
- Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, 25(1), 23–38. <https://doi.org/10.1080/02602930050025006>
- Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of Control. *Applied Linguistics*, 14(1), 25–42. <https://doi.org/10.1093/applin/14.1.25>
- Ploegh, K. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation*, 35(2–3), 102–109. <https://doi.org/10.1016/j.stue-duc.2009.05.001>
- Runnels, J. (2014). Japanese English learner self-assessments on the CEFR-J's A-level can-do statements using four and five-point response scales. *The Asian Journal of Applied Linguistics*, 1(2), 167–177. <https://caes.hku.hk/ajal/index.php/ajal/article/view/13>
- Sadler, P. R., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, 11(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1
- Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentation. *RELC Journal*, 40(2), 149–171. <https://doi.org/10.1177%2F0033688209105868>
- Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research*, 68(3), 249–276. <https://doi.org/10.3102/00346543068003249>
- Tsui, A., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing*, 9(2), 147–170. [https://doi.org/10.1016/S1060-3743\(00\)00022-9](https://doi.org/10.1016/S1060-3743(00)00022-9)
- Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255–274). De Gruyter. <https://doi.org/10.1515/9781614513827-018>
- van Gennip, N. A., Segers, M. S. S., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction*, 20(4), 280–290. <https://doi.org/10.1016/j.learninstruc.2009.08.010>

- Wiliam, D. (2000). *Integrating summative and formative functions of assessment* [Keynote address]. European Association for Educational Assessment, Prague, Czech Republic. https://discovery.ucl.ac.uk/id/eprint/1507176/1/Wiliam2000IntegratingAEA-E_2000_keynoteaddress.pdf
- Zarei, A. A., & Mahdavi, A. (2014). The effect of peer and teacher assessment on EFL learners' grammatical and lexical writing accuracy. *Journal of Social Issues & Humanities*, 2(9), 92-97.
- Zarei, A. A., & Usefli, Z. (2015). The effect of assessment type on EFL learners' goal orientation. *Journal of Language, Linguistics and Literature*, 1(4), 112-119. <http://creativecommons.org/licenses/by-nc/4.0/>

付録 A

Writing Assessment Rubric

Components	Assessment criteria	Scores								
Task Fulfillment (課題で求められている内容が含まれているかどうか)	Does English composition fulfill the task? - present one's idea or opinion - present two reasons -present concrete explanation and examples 課題で求められている内容が含まれていますか? 自分の考えや意見を明示していますか。 その理由が二つ書かれていますか。 理由には具体的な例や説明が加えられていますか。	<table><tr><td colspan="2">Good</td><td colspan="2">Poor</td></tr><tr><td>4</td><td>3</td><td>2</td><td>1</td></tr></table>	Good		Poor		4	3	2	1
Good		Poor								
4	3	2	1							
Structure and Coherence (構成と内容の一貫性)	Is composition written coherently and logically? -easy to understand -use discourse markers effectively -well organized 英作文は内容に一貫性があり，論理的ですか。 内容や表現に一貫性があり，読んでいてわかりやすいですか。 伝えたい情報の流れや展開を示す表現(接続詞など)を効果的に使っていますか。 関係のない考えや情報が書かれていず，まとまりがありますか。	<table><tr><td colspan="2">Good</td><td colspan="2">Poor</td></tr><tr><td>4</td><td>3</td><td>2</td><td>1</td></tr></table>	Good		Poor		4	3	2	1
Good		Poor								
4	3	2	1							
Appropriate Usage of Vocabulary (語彙の適切な使用)	Is vocabulary used appropriately and accurately? -accurate spelling -appropriate usage of vocabulary 語彙は適切で正確に使われていますか。 単語の綴りは正確ですか。 単語は正しい意味で使われていますか。 英語以外の言葉を使う場合は，その言葉を知らない人でも理解できるように説明が加えられていますか。	<table><tr><td colspan="2">Good</td><td colspan="2">Poor</td></tr><tr><td>4</td><td>3</td><td>2</td><td>1</td></tr></table>	Good		Poor		4	3	2	1
Good		Poor								
4	3	2	1							
Grammatical Accuracy (正確な文法の使用)	Is grammar used to effectively convey ideas and reasons? 自分の考えとその理由を効果的に伝えられるように文法は効果的に使われていますか。	<table><tr><td colspan="2">Good</td><td colspan="2">Poor</td></tr><tr><td>4</td><td>3</td><td>2</td><td>1</td></tr></table>	Good		Poor		4	3	2	1
Good		Poor								
4	3	2	1							
Sum(合計)		/16								

*次に全体をふりかえって自分の英作文の評価を日本語か英語でコメントしてください。

My English writing is _____.

私(ぼく)の英作文は全体的に_____。

NAME ()

付録 B

英作文課題一覧

	自己評価グループ	ピア評価グループ
実験前のライティング テスト	What is your favourite season?	
(評価練習)	Which country do you want to visit?	
1st session	What do you do in your free time?	
2nd session	Where would you like to go in this summer holiday?	
3rd session	Which do you like better, staying at home or spending time out of the house?	
4th session	What is your favourite food?	
実験後のライティング テスト	What is your future dream?	