第37回 研究助成

A 研究部門 ● 報告Ⅲ ● 英語能力テストに関する研究

生成 AI を用いたルーブリック評価の妥当性と可能性の探究:教員評価と生成 AI による評価の比較研究

研究者:大和田 彩　高知県／高知県立室戸高等学校　教諭(申請時:米国／ハワイ大学マノア校大学院　在籍)
《研究助言者:竹内 理》

概要

本研究の目的は,高校生の英作文に対する従来の教員による評価(以下,教員評価)と,生成AI(ChatGPT)による評価結果を比較し,生成AIを活用したルーブリック評価の妥当性と可能性を検証することである。近年,生成AIを教育評価に活用することで,評価の効率化や公平性の向上が期待されているが,その評価結果が教員評価とどの程度一致するのか,またAI評価に特有の傾向が存在するのかについては,十分に検証されていない。本研究では,高校生30名の英作文を対象に,教員20名による評価とChatGPTの評価を比較した。分析の結果,文法・語彙・構成といった形式的観点ではAIと教員の評価が高い一致度を示すことが確認された一方で,主体的態度といった主観的観点では,AIは教員よりも過小評価する傾向がみられた。さらに,Technology Acceptance Model(TAM)に基づくアンケートでは,教員はAIの有用性や効率性を一定程度認めつつも,「主観的観点の判断は人間が担うべき」とする慎重な姿勢が示された。

1 はじめに

英作文は,英語教育において学習者の論理的思考力や表現力,さらには批判的思考力を育成するための重要な学習活動のひとつである。単なる語彙や文法の習得にとどまらず,自分の考えを筋道立てて表現し,相手に伝えるという実践的な力を育てる場として,英作文は中等教育から高等教育にかけて幅広く導入されている。その一方で,評価の在り方については長年にわたり課題が指摘されてきた。Bloom et al.(1956)の教育目標分類で示されているように,英作文の評価には知識や理解だけでなく,応用,分析,統合といった高次の認知的プロセスが関与しており,単純な正誤判定では捉えきれない複雑な判断が求められる。従来の英作文評価は,教員の専門的判断に基づいて行われることが多く,その専門性は評価の質を担保する一方で,評価者間の一貫性や公平性を確保するうえでの大きな障壁ともなってきた。特に,大規模なクラス編成や多忙な学校現場においては,すべての生徒に対して十分な時間を割いた丁寧な評価とフィードバックを行うことは容易ではない。Ellis et al.(2006)は,こうした教育実践において,リキャストのような暗示的フィードバックよりも,メタ言語的説明を伴う明示的フィードバックの方が学習者の理解を深めやすいと指摘している。中でも,ルーブリックやコメントを活用した訂正フィードバック(Written Corrective Feedback:WCF)は,誤りの明示と修正を通じて言語習得を促進する有効な手法とされ,実際の授業においても高く評価されている(Zeevy-Solovey, 2024)。しかしながら,すべての教員がこのような包括的なフィードバックを実践するには,人的,時間的リソースの限界があるのが現実である。大人数のクラスで一人ひとりの作文を丁寧に読み,誤りを指摘し,改善の方向性を提示する作業は,多く

の教員にとって大きな負担となっている(Lin & Crosthwaite, 2024; Shermis et al. 2002)。このような状況において, 学習者にとっても, 必要なタイミングで十分なフィードバックを受け取る機会が限られてしまうという課題が生じている。

こうした背景を受けて, 近年では教育現場における評価支援ツールとしてのAI技術の活用が注目されている。とりわけ, 言語学習支援を目的とした生成AIの進展は目覚ましく, 自然言語処理能力の向上により, 質の高いフィードバックを瞬時に提供することが可能となりつつある(ElEbyary et al., 2024)。生成AIは, 学習者の英作文に対して個別的かつ即時的な応答を生成し, 学習者の理解を深め, 主体的な改善行動を促すツールとしての可能性を秘めている。文部科学省(2024)も, 生成AIを「伴走者」として位置づけ, 教員の負担軽減と学習者の主体性育成の両立を図る補助的な活用を推奨しており, その導入に対する社会的認知も進みつつある。さらに, Technology Acceptance Model(TAM)の理論的枠組みによれば, 新しい技術が教育現場に受け入れられるかどうかは, 有用性と使いやすさによって大きく左右されるとされている(Davis, 1989; Chocarro et al., 2023)。つまり, どれほど技術的に優れたツールであっても, 教員や学習者がその利便性や必要性を実感できなければ, 教育実践への定着は困難である。したがって, AI導入に際しては, 技術的条件のみならず, 教育関係者の心理的な受容性に配慮した支援や研修の整備が不可欠であるといえる。

以上を踏まえ, 本研究では, 日本の高等学校における英作文指導の実態に即して, 高校生による英作文に対する教員評価と, 生成AI(ChatGPT)による評価結果を比較し, AIを活用した英作文評価の妥当性と有効性を検討する。また, 教員および学習者のAI評価に対する受容意識を明らかにすることで, 今後の教育現場におけるAI導入に向けた具体的かつ実践的な示唆を提示することを目的とする。

2 先行研究

2.1 L2ライティングにおけるフィードバックの役割と課題

第二言語習得(Second Language Acquisition: SLA)に関する研究領域では, 学習者の言語運用能力や習得プロセスを説明するために, さまざまな理論的枠組みが提案されてきた。その中でも, 行動主義の立場に基づく理論では,「すべての行動は刺激と反応による条件づけの結果である(VanPatten & Williams, 2007)」とされ, 学習はさまざまなフィードバックの積み重ねによって形成されると理解されている。この観点は, 英作文指導におけるフィードバックの意義と深く関係しており, 学習者の言語表現力を高めるための要因として, 教員のフィードバックが重要な役割を担っている。フィードバックは, 学習者に対して自身のエラーや課題を気づかせ, 改善へと導くための具体的な情報を提供するものであり, 英作文の完成度を高めるための中心的手段である(Ferris & Roberts, 2001; Troia, 2014)。特に, エラーの種類や学習段階に応じた適切なフィードバックの方法が, 学習者の言語習得に与える影響は大きい。Ellis et al.(2006)は, さまざまなフィードバックの形式の中でも, 明示的な訂正フィードバックが学習者の注意を引きやすく, エラー修正の定着において高い効果を示すことを明らかにしている。さらに, Teimouri(2017)は, 単なる訂正にとどまらず, 学習者の努力や長所に光を当てるポジティブフィードバックを併用することが, 学習意欲や自信の向上につながると指摘しており, 感情的側面への配慮の重要性を示唆している。加えて, Troia(2014)は, フィードバックの在り方について, 教員から学習者への一方向的なコメントに限定するのではなく, 学習者との双方向的なコミュニケーションを通じて意味交渉を行う「対話的フィードバック(dialogic feedback)」の有効性に注目すべきであると述べている。こうしたフィードバックは, 単なる誤りの指摘ではなく, 学習者が自らの言語表現を見直し, 再構成する過程に主体的に関わる機会を提供する点で, 従来の指導よりも深い学びを促すことが期待されている。

しかしながら, このように理想的なフィードバックを実現するためには, 教員に多くの時間的, 精神的

負担がかかるのが現実である。特に，クラスの規模が大きく，一人の教員が複数の英作文に対応しなければならない場合，公平性や一貫性を維持しながら個別にフィードバックを行うことは非常に困難である。また，学習者によって必要とする支援の内容が異なるため，画一的な指導では対応しきれないというジレンマも存在する。このような状況を踏まえると，英作文における効果的なフィードバックを持続的に実施するためには，教員を支える補助的な手段の導入が不可欠であるといえる。近年では，教育技術の進展に伴い，AIをはじめとするテクノロジーを活用した支援ツールが注目を集めており，こうした技術をどのように教育実践に統合し，学習効果の向上につなげていくかが，今後の大きな課題となっている。フィードバックの質と量の両立，学習者の個別性への対応，教員の負担軽減といった複合的な課題に対し，テクノロジーの活用がいかなる可能性を持つのかを検証することは，今後の英作文指導の実践にとって極めて重要な視点である。

2.2 AIによるライティング評価の可能性と課題

近年，生成AIの急速な技術革新により，言語教育への応用が著しく進展している。従来のAWE（自動作文評価）システムに比べて，ChatGPTのような生成AIは柔軟かつ学習者にとって理解しやすいフィードバックを提示できる点が注目されており（Han & Li, 2024），特にその即時性と一貫性は，正確かつ効率的なフィードバックの提供に貢献する（Polakova & Ivenz, 2024; Mohammed & Khalid, 2025）。一方で，Teng（2024）の研究によれば，EFL学習者はChatGPTのフィードバックの有効性を認めつつも，人間味や具体性の欠如などの限界を指摘しており，教員による補完の必要性が示唆されている。

AIの評価的活用に関しては，その信頼性と妥当性が実用化に向けた重要な課題とされている。Yang（2024）は，ChatGPTにおける自己内一致信頼性の低さを指摘し，同一基準に基づく安定した評価の困難さを示している。また，Uyar and Büyükahıska（2025）は，AIと人間評価者によるスコアに有意差が見られ，特に創造性や論理構成といった高次スキルにおいてその差が顕著であることを報告している。Yao et al.（2025）は，生成AIを補助ツールとしてL2ライティング指導に組み込む際，教員がどのように活用するか，その実践的統合のプロセスに不確実性があると指摘しており，Yavuz et al.（2025）も，LLM（大規模言語モデル）の出力に過度に依存することで，評価の信頼性が損なわれる可能性を示している。さらに，Steiss et al.（2024）はAIのフィードバック品質は一定水準を維持しているとする一方で，Fuller and Bixby（2024）は，生成AIの使用方法によっては採点やフィードバックにばらつきが生じ，信頼性や一貫性に欠ける可能性があると指摘している。

このように，AI評価は即時性や一貫性といった利点を有し，学習者のアウトプットの改善や教員の負担軽減に貢献し得るが，最終的な評価は教員が担うべきであり，AIは補助的手段として活用されるべきである（Li et al., 2024）。その精度を高めるには，明確なルーブリックと評価者間の信頼性（Inter Rater Reliability: IRR）の確保が不可欠である（Kaldaras et al., 2022）。ルーブリックは主にパフォーマンスの評価に活用される（Brookhart, 2013）。また，AIの導入に際しては，教育現場における倫理的配慮や人間中心の価値観を損なわない戦略的運用が求められる（Pozdniakov et al., 2024）。教員によるフィードバックは誤りの特定や内省的思考の促進に有効であり，AIは低次の誤りへの対応に強みを持つ（Fan et al., 2024）。ChatGPTの統合型ライティングタスクへの有効性については今後の検証が必要とされ（Kim et al., 2024），AIと人間のフィードバック併用による効果もまだ十分に研究されていない（Tran, 2025）。Tate et al.（2024）の分析では，AI採点による明確な悪影響は認められず，プロンプトやコーパスの工夫によって改善の余地があることが示された。したがって，AIと人間による評価の差異および相補的可能性を明らかにすることは，教育現場における効果的かつ持続可能な評価の在り方を再考するうえで，重要な意義を持つ。

2.3 AI評価の受容と課題

AIを教育現場に導入する際には，単に技術的な精度や機能性の高さのみならず，それを使用する教員や学習者自身の「受容意識」が極めて重要な要素として作用する。つまり，技術が有しているポテンシャルが教育実践において十分に活かされるかどうかは，実際に利用する人々の態度や意識によって大きく左右されるのである。この点について理論的枠組みを提供するのが，Davis（1989）によって提唱された技術受容モデル（Technology Acceptance Model: TAM）である。TAMは，新たな技術が利用者に受け入れられるかどうかを予測するための代表的なモデルであり，「知覚された有用性（Perceived Usefulness: PU）」および「知覚された使いやすさ（Perceived Ease of Use: PEOU）」，「行動意図（Behavioral Intention: BI）」という三つの主要な概念に着目する。ここで言う「有用性」とは，あるシステムやツールを使用することで自身の仕事や学習のパフォーマンスが向上すると利用者が信じる程度を指し，「使いやすさ」とは，そのシステムの操作が困難ではなく，ストレスなく活用できると感じる程度を意味する。これら二つの知覚が，利用者の態度や使用意図（BI）を大きく左右し，最終的な利用行動に結び付くとされている。このモデルは，ICT機器やAIツールといった教育テクノロジーを学校教育へ導入する際にも，極めて有効な理論的枠組みとして多くの研究に採用されている。

さらに，近年の研究ではTAMを発展的に活用し，実際の教育現場におけるAI導入の受容要因についての実証的分析が進められている。例えば，Chocarro et al.（2023）の研究は，教員の年齢や一般的なデジタルスキルよりも，AIツールの「使いやすさ」および「有用性」に対する主観的な認知の方が，AIの活用意図を高めるうえでより強力な予測因子であることを明らかにした。つまり，教員が若いかどうかやICTに熟達しているかよりも，「このツールは役に立ちそうだ」「簡単に使えそうだ」と感じられることが，AI受容の鍵を握っているという指摘である。また，Lai et al.（2023）は，学習者側の動機づけタイプとAIツールの受容との関係に注目し，学習者の内発的，外発的動機づけの違いが，AIの有用性認知や使用意図に異なる形で影響を与えることを示した。具体的には，外発的動機づけ（成績や報酬など外的要因による動機）が強い学習者は，AIの「有用性」に対する評価が使用意図に影響を及ぼしやすく，一方で，内発的動機づけ（知的好奇心や自己成長への欲求）が高い学習者は，「使いやすさ」の認知が使用意図により強く関与することが確認されている。これにより，AIツールを導入する際には，学習者の動機づけの傾向に応じてアプローチを工夫する必要があることが示唆される。

先にも述べたように，文部科学省（2024）は生成AIを教育現場における「伴走者」として位置づけており，教員の業務負担を軽減する一方で，学習者の主体的な学びを支援することが望ましい活用の在り方としている。しかし，このような理想的な活用を実現するためには，教員や学習者がAIによる評価に対して十分な信頼を寄せ，安心して活用できる心理的環境が整っていることが前提条件となる。

3 研究，実践，調査方法

3.1 調査の目的

本研究は，生成AI（ChatGPT）による英作文評価と高等学校英語教員による評価を比較し，その一致度や評価傾向を明らかにするとともに，AI評価の教育的活用の可能性を検討することを目的とする。具体的には，まずAIと教員がどの程度同じ判断を下しているのかを確認する（RQ1）。次に，その一致度を補足的に理解するため，観点別にAIが一貫して過小または過大評価を行う傾向が存在するのかを分析する（RQ2）。これら二つの問いによって，AI評価の特徴を明らかにする。そのうえで教育現場におけるAI評価の導入の可能性を探るために，教員がAI評価をどのように受容し，どの場面で補助的に活用できると考えているのかを検討する（RQ3）。以上を踏まえ，以下の3点を検証課題（Research Questions: RQ）と

して設定した。

RQ1 生成AIによる英作文評価は，教員による評価とどの程度一致するのか。

RQ2 評価観点ごとに，生成AIに特有の過小または過大評価の傾向は見られるか。

RQ3 RQ1と2で明らかになった一致度や傾向差を前提として，高等学校英語教員は生成AIによる英作文評価をどのように受容しているか。

3.2 研究デザインと評価実施手順

本研究は，日本国内の高等学校に勤務する英語教員による英作文評価と，生成AI（ChatGPT）による観点別評価とを比較することで，それぞれの評価傾向，安定性，実用性について実証的に検討し，AIの補助的活用の可能性と教員の受容要因を明らかにすることを目的として実施された。

評価対象となった英作文は，筆者が勤務する高等学校に在籍する，研究実施当時高校2･3年生の30名が書いたものである。英語力については，CEFRで概ねA2～B1レベルに相当している。テーマは「あなたが学びたい第三言語とその理由」とし，主張と二つの理由を明確に示す構成で書くように指導を行った。語数の目安は100～120語とし，オンラインを含む辞書の使用は許可したが，翻訳ツールや文法補助機能などの使用は禁止した。提出された英作文はすべて無記名化され，評価用データセットとして整理，保存された。

評価には，英検の採点基準に準じたルーブリックと，新学習指導要領に準拠した観点別ルーブリックの2種類を使用した。英検準拠型では「内容」「構成」「語彙」「文法」の4観点について，それぞれ0～4点の5段階で評価を行い，学習指導要領準拠型では「知識，技能」「思考，判断，表現」「学習に主体的に取り組む態度」の3観点について，a（優れている）～c（改善が必要）の3段階で評価を行った。教員による評価は，全国の高等学校に勤務する英語教員20名を対象として実施した。年齢層は20代から50代まで幅広く，教員経験年数も1年未満から11年以上と多様であった。評価に用いるルーブリックは学習指導要領準拠，学校独自の様式，個人作成，英検基準などさまざまであり，ICTや生成AIの活用度も積極的な者から未経験の者まで幅広い分布を示した。各教員には30作品のうち6作品を無作為に割り当て，両方のルーブリックに基づいて観点別に評価してもらった。事前に評価マニュアルを配布し，評価者間の一貫性を担保するよう努めた。すべての評価はGoogleフォームを用いてスコア形式で回収し，自由記述によるフィードバックや文法などの誤りについての修正提案などは求めなかった。一方，AIによる評価には，有料版のChatGPT（GPT-4o）を使用した。それぞれのルーブリックに対応する2種類のカスタムGPT（英検準拠型および学習指導要領準拠型）を作成し，同一の30作品に対して観点別のスコアを出力させた。各英作文については複数回（最低3回，最高5回）評価を生成し，出力の整合性を確認したうえで，内容が一致した場合にその結果を最終スコアとして採用した。AIによる評価結果はすべてチャットGPT履歴のURLとともに記録，保存され，検証可能な形で管理された。収集された評価結果のうち，複数のスコアが併記されていたものなど一意に解釈できないデータは「分析不能データ」として除外し，最終的に教員評価113件，AI評価120件を分析対象とした。

3.3 評価結果の分析

本研究では，AIと教員による評価結果の傾向や差異，そして評価の一貫性や安定性の比較を通じて，生成AIによる英作文評価の有効性と限界を検証することを目的とした。まず，両者の評価スコアの平均値に有意な差が存在するかを明らかにするため，各観点において対応のある t 検定（Paired-Samples t-Test）を実施し，その差の大きさを定量的に把握するために効果量（Cohen's d）を算出した。次に，AI評価と教員評価の傾向の違いを視覚的に把握するため，各観点における「教員評価－AI評価」の差分スコア

を算出し，箱ひげ図（boxplot）によってその分布を可視化することで，AIが特定の観点において一貫して過小または過大評価を行う傾向の有無を検証した。加えて，AIと教員がそれぞれどの程度安定した評価を行っているかを確認するため，英作文ごとに各観点スコアの標準偏差を算出し，評価のばらつきを比較した。さらに，両者の観点別スコアや平均値がどの程度連動して変化するかを確認するため，Pearson's *r* を用いて相関分析を行った。これらすべての統計分析はSPSSを用いて実施した。

3.4 AIの活用に対する意識調査

AIによる英作文評価に対する教員の受容意識を把握するため，評価実施の前後にアンケート調査を実施した。調査設計にあたっては，Davis（1989）のTechnology Acceptance Model（TAM）を理論的枠組みとして採用し，「知覚された有用性（Perceived Usefulness）」「知覚された使いやすさ（Perceived Ease of Use）」「行動意図（Behavioral Intention：BI）」「利用態度（Attitude Toward Using：ATT）」「社会的影響（Social Influence：SI）」の5構成要素に対応する計32項目を作成した。すべての設問は5段階のリッカート尺度により構成されており，回答者の主観的評価を数量化できるよう設計されている。事前アンケートでは，年齢層，指導年数，ICT活用頻度，AIの使用経験，AIへの期待感や抵抗感など，教員の基本属性および初期意識を収集した。事後アンケートでは，実際のAI評価体験を踏まえた印象や今後の導入意向を把握することを目的とした（資料1）。ただし，本研究では教員の一般的な受容傾向を明らかにすることに焦点を当てた。また，自由記述では生成AIを使用する際の懸念と期待を尋ね，TAMに基づく分析結果を考察する際の材料とした。

4 結果および考察

4.1 AI評価，教員評価の一致度分析（対応のあるサンプルの*t*検定）

英検準拠ルーブリック（内容，構成，語彙，文法）および新学習指導要領準拠ルーブリック（知識，表現，主体的態度）の各観点において，AI評価と教員評価との一致度を検証するため，対応のあるサンプルの*t*検定を実施した。さらに，差の大きさを定量的に把握するために効果量（Cohen's *d*）を算出し，その解釈には Plonsky and Oswald（2014）の基準を用いた。加えて，両者の関連性を確認するために Pearson's *r* を算出した。結果の概要を表1および表2に示す。

英検準拠ルーブリック（表1）については四つの観点のうち，「内容」に関しては AI と教員の評価の平均がほぼ同一であり，有意な差は確認されなかった。また，相関係数は非常に高く，AIによる判断が教員とおおむね一致していることが示された。一方，「構成」「語彙」「文法」においては，いずれも AI 評価の方が教員評価よりもやや低く，統計的に有意な差が認められた。特に「文法」では，効果量 $d = -0.58$ が中程度に該当し，AI が文法面を過小評価する傾向が示唆された。これに対し，「構成」「語彙」の効果量は 小程度にとどまり，実質的な差は限定的であった。ただし，これらの観点においても中〜高程度の相関が確認され，評価の傾向は基本的に整合していたといえる。

学習指導要領準拠ルーブリック（表2）の3観点では，「知識」と「主体的態度」において，AI の評価が教員より有意に低いことが明らかとなった。特に「知識」の効果量 $d = -0.80$は 大きい水準に該当し，AI と教員の認識の違いが顕著であった。また，「主体的態度」の効果量 $d = -0.67$は 中程度に位置づけられ，この観点でも AI 評価の限界が浮き彫りとなった。一方，「表現」については平均値の差は小さく，有意な差は認められなかった。効果量 $d = -0.18$ は 小程度であり，実質的な差はほとんどないと解釈できる。これら3観点すべてにおいて，中〜高程度の正の相関が確認され，AI 評価は全体として教員評価とよく一致していると考えられる。

以上のことから，AI評価と教員評価は観点ごとに一定の一貫性を示し，高い相関関係が確認された。しかし，英検準拠ルーブリックにおける「文法」や，新学習指導要領準拠ルーブリックにおける「知識」「主体的態度」では，AIが一貫して教員よりも厳しい評価を行う傾向が明らかとなった。これらの結果は，AI評価の有効性を支持する一方で，特定観点における過小評価の可能性にも留意する必要があることを示している。

■表1: 英検準拠ルーブリックにおけるAI評価と教員評価の比較（*N* = 30）

観点	教員平均（SD）	AI平均（SD）	*t*値	*p*値	効果量（*d*）	相関係数（*r*）
内容	3.06 (.80)	3.01 (.82)	-0.46	.650	-0.08	.795
構成	2.97 (.70)	2.76 (.81)	-2.19	.037	-0.40	.765
語彙	3.05 (.72)	2.85 (.76)	-2.31	.028	-0.42	.722
文法	3.22 (.65)	2.93 (.72)	-3.16	.004	-0.58	.807

■表2: 学習指導要領準拠ルーブリックにおけるAI評価と教員評価の比較（*N* = 30）

観点	教員平均（SD）	AI平均（SD）	*t*値	*p*値	効果量（*d*）	相関係数（*r*）
知識	2.11 (.55)	1.80 (.52)	-4.38	<.001	-0.80	.743
表現	2.01 (.48)	1.93 (.51)	-1.00	.325	-0.18	.716
主体的態度	2.12 (.50)	1.85 (.53)	-3.66	<.001	-0.67	.784

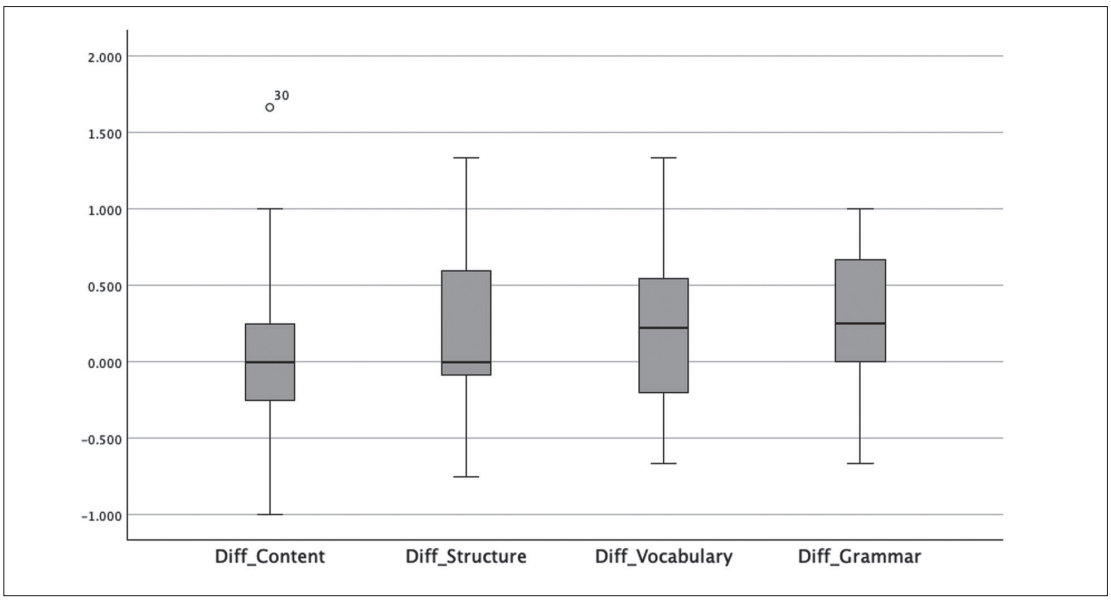
4.2 AI評価の傾向分析（箱ひげ図）

AIと教員の評価傾向の違いを視覚的に捉えるため，各観点における「教員評価－AI評価」の差分スコアを算出し，その分布を箱ひげ図で示した。ここで正の値は教員評価の方がAI評価を上回っていること，負の値はAI評価が高かったことを意味する。本研究では，英検準拠の観点（内容，構成，語彙，文法）および学習指導要領準拠の観点（知識，表現，主体的態度）に分けて分析を行い，それぞれ図1および図2に示した。

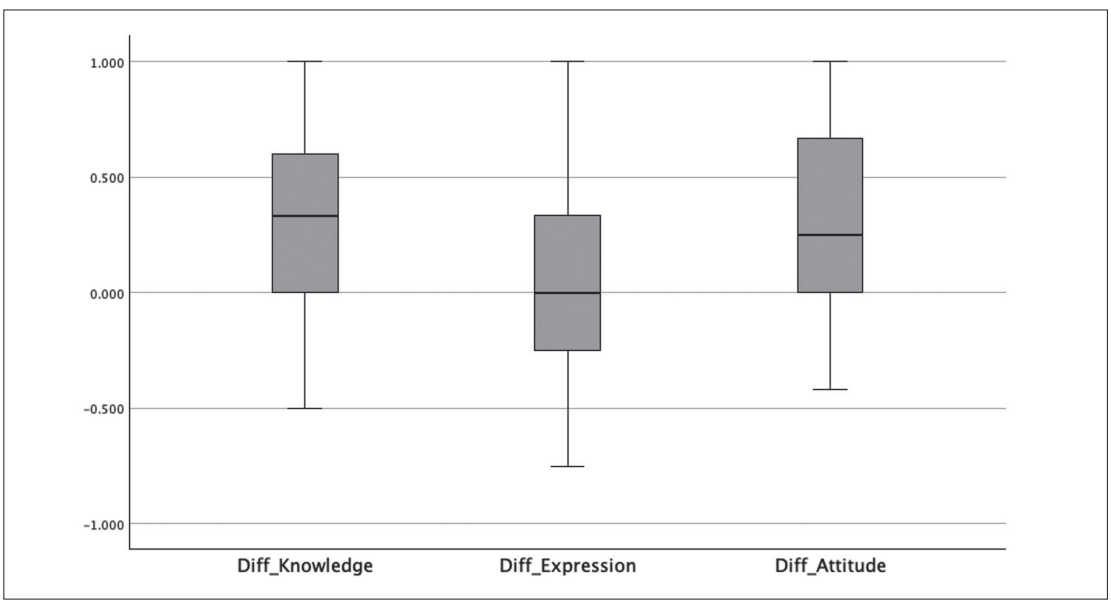
英検準拠ルーブリックについては（図1），4観点すべてにおいてAI評価が教員評価を下回る傾向が明確に示された。特に「文法」においては，差分スコアの中央値が他の観点よりも大きく，AIがこの領域で一貫して厳しめの評価を行っていることがうかがえる。また，「語彙」や「構成」においても，教員評価の方がやや高い傾向が確認されており，いずれも評価差のばらつきは比較的少なく，AIの判断が一定の傾向に収束している様子が見て取れる。

学習指導要領準拠ルーブリックについては（図2），特に「知識」および「主体的態度」において，教員評価がAI評価を明確に上回っており，差分の中央値はおおよそ +0.3〜+0.5 の範囲に分布していた（図2）。内容（Content）の観点では作文30番に外れ値が確認され，AIが当該作文に著しく低いスコアを付与したことが示された。これは，AIが主観的な要素や学習態度のような内面的な観点を適切に判断することに限界がある可能性を示している。一方，「表現」については，中央値が0に近く，ばらつきも比較的小さかったことから，論理構成やアイデア展開といった比較的客観的な評価項目においては，AIと教員の判断が一致しやすいことがうかがえる。

以上の結果から，AIは「文法」や「構成」など明示的な形式的基準に対しては安定的かつ妥当な判断を下す一方，「知識」や「主体的態度」といった学習内容の正確さや学習への姿勢など，文脈や内面の理解が求められる観点では，AIが評価を低くつける傾向が見られた。特に「文法」「知識」「態度」に見られた評価差は，AI評価の限界を示唆するものであり，今後の活用に際しては，観点ごとの特性やバイアスを考慮し，教員による補完的な判断との併用が望ましいといえる。



■図1：英検準拠ルーブリックにおける教員評価 − AI評価の差（箱ひげ図）



■図2：新学習指導要領準拠ルーブリックにおける教員評価 − AI評価の差（箱ひげ図 ）

4.3 評価の一貫性分析（Intraclass Correlation Coefficient; ICC）

本研究では，教員および生成AIによる英作文評価の信頼性と一貫性を明らかにするため，各観点において級内相関係数（Intraclass Correlation Coefficient: ICC）を用いて，評価の一致度を比較した。ICCは，同じ英作文に対して複数の評価者がどの程度似た評価をしているかを数値で示すもので，教育評価の分野でもよく使われている指標である（Shrout & Fleiss, 1979）。分析にあたっては，AIと教員の両方について，「Two-Way Random, Absolute Agreement, Average Measures」と呼ばれる共通の方法を用いてICCを算出した。これは，評価する人と英作文の両方をランダムな要素とみなして，どれくらい評価が一致しているかを確かめる方法である。AIも教員も同じ方法で分析したことで，どちらの評価がど

の観点で安定していたのかが明確になった(表3)。例えば, 文法の観点ではAIの評価が非常に安定しており(ICC = .919), 一方で「知識」や「主体的態度」など, 内容の理解が必要な観点では, 教員評価の方が一貫している傾向も見られた(表4)。ICCの数値の見方としては, Cicchetti(1994)の基準を参考にし, .75以上を「非常に良い(excellent)」, .60〜.74を「良い(good)」, .40〜.59を「中程度(fair)」, .40未満を「信頼性が低い(poor)」として分析を行った。

■表3: 英検準拠ルーブリックにおける教員評価とAI評価のICC比較

観点	教員評価 ICC	AI評価 ICC	評価傾向の比較
内容	.650	.556	教員, AIともに fair 程度の一貫性
構成	.747	.885	教員:good, AI:excellent
語彙	.915	.880	両者とも excellent, 一貫性が高い
文法	.703	.919	教員:good, AI:excellent

表3に示すとおり, 語彙(Vocabulary)においては教員, AIいずれも非常に高いICC値を示し, .75以上の「非常に良い(excellent)」水準に分類された。特に, 両者とも.88を超える値を示しており, 既存の基準における「excellent」の中でも際立って高い信頼性が確認された。このことから, 語彙使用の適切さや多様性といった観点は, 比較的客観的かつ明確な基準で評価されていることが示唆される。構成(Structure)および文法(Grammar)においても, AI評価はそれぞれ .885, .919 と非常に高いICCを示し, 教員評価(.747, .703)を上回った。これは, AIが構成力や文法的正確性に関して一貫した評価基準を保持していることを示している。一方, 内容(Content)の観点では, 教員評価が .650, AI評価が .556 にとどまり, いずれも高い一貫性とは言えず, 教員, AIともに fair レベルにとどまった。また, 教員評価においては, 95%信頼区間にマイナス値(CI: -0.210〜0.959)が含まれており, 統計的に一貫性が確認されたとは言いがたい結果であった。これは, 主張の明確さや説得力といった定性的, 主観的要素が強く, 評価者間で判断が分かれやすい観点であることを示唆する。AIにおいても, 意味解釈を要する観点では一貫性のある処理が難しい。

■表4: 学習指導要領準拠ルーブリックにおける教員評価とAI評価のICC比較

観点	教員評価 ICC	AI評価 ICC	評価傾向の比較
知識	.932	.471	教員:excellent, AI:fair
表現	.721	.900	教員:good, AI:excellent
主体的態度	.688	.200	教員:good, AI:poor

学習指導要領準拠の観点においては(表4), 知識(Knowledge)で教員評価は.932と極めて高く, 観点の定義や評価基準が教員間で共有されていたことが示される。一方で, AI評価は.471にとどまり, fair レベルであり, 知識の適切さや情報の正確性の判断において一貫性が保たれていないことが示された。これは, 文脈理解や背景知識の活用を必要とする観点において, 生成AIの処理能力が限定的である可能性を示している。表現(Expression)については, AIが.900とexcellent レベルの高い一貫性を示し, 教員評価(.721)よりも高い水準であった。これは, AIが論理構成, 段落展開, 接続表現などの形式的な構造要素に関して, 安定した基準で評価をしていることを意味している。特に, 構文や形式的特徴に強みを持つAIの特性が, この観点において顕著に発揮されたと考えられる。対照的に, 主体的態度(Attitude)の観点では, 教員評価でも.688とやや低めであり, AI評価は.200とpoorレベルに分類された。この結果は, 学習への意欲や内面的な姿勢といった主観性の高い観点において, AIが一貫性をもって評価することが極めて難しいという実態を示している。

4.4 教員の受容意識（TAM分析）

本研究では，生成AIを用いた英作文評価の有効性を検討するにあたり，その導入，活用に対する高等学校英語教員の受容意識を把握することを目的として，TAMに基づいた事前・事後アンケート調査を実施した（資料1）。調査対象は，本研究に協力した高等学校英語教員20名であり，TAMの五つの主要構成要素に対応する計32項目について，5段階リッカート尺度による評価を依頼した。構成要素は以下のとおりである：知覚された有用性（PU），知覚された使いやすさ（PEOU），行動意図（BI），使用態度（ATT），および社会的影響（SI）。質問項目の詳細は資料1「教員の受容意識（TAM分析）アンケート調査項目」に示した。分析においては各変数ごとに平均値を算出し，次に行動意図を目的変数とした重回帰分析（multiple regression analysis）を実施し，PU, PEOU, ATT, SIがBIにどのような影響を与えているかを検討した（表5）。分析の結果，「知覚された有用性」は「行動意図」に統計的に有意な負の影響を与えていた（$p < .05$）（$\beta = -0.651$）。これはTAMの理論的前提とは異なるものであった。本来，PUが高いほどBIも高くなる（＝便利だと思えば使いたくなる）とされているが，今回のデータでは「便利だと感じても，使いたくない」という教員の認識が浮き彫りになった。この逆転現象の背景には，アンケートの自由記述で述べられた次のような意見が関係していると考えられる：

「AIで評価しても結局妥当性を確認するために読み直す必要があり，手間が増えるだけ」「プロンプトによって結果が大きく変わるのは不安」「生徒のことをきちんと理解するには，やはり自分で英作文を読む必要がある」つまり，教員はAI評価の利便性や機能面を理解していても，それが現場での実践に直結しないと感じている可能性がある。評価に対する責任感や，生徒との教育的関係を重視する姿勢が，AIの導入を躊躇する要因となっているのかもしれない。また，「便利だがプロンプト次第で結果が変わる」という意見に見られるように，AIの安定性や妥当性に対する不安感が，知覚された有用性（PU）を逆方向に作用させているとも考えられる。

■表5: BI（行動意図）に対する各要因の重回帰分析結果（n=20）

変数	標準化係数（β）	t値	有意確率（p値）	有意性
知覚された有用性	-0.651	-2.558	.022	有意（p<.05）
知覚された使いやすさ	0.254	1.103	.287	有意でない
利用態度	0.208	0.707	.490	有意でない
社会的影響	0.320	1.161	.264	有意でない

4.5 考察

本研究では，生成AIによる高校生英作文のルーブリック評価の妥当性，AI評価が実際の教育実践にどのように受け入れられるのかを検証することを目的とした。以下では，本研究で設定した研究質問（RQ）に基づき，得られた結果を踏まえて考察を行う。

RQ1：生成AIによる英作文評価は，教員による評価とどの程度一致するのか。

➡観点によって一致度に差があることが明らかになった。

対応のあるt検定および相関係数の結果，「文法（Grammar）」「語彙（Vocabulary）」「構成（Structure）」といった形式的・明示的な観点では，AIと教員の評価は平均値の差が小さく，相関係数も高かったことから，おおむね一致していることが示された。一方，学習指導要領準拠によるルーブリック評価における主体的態度（Attitude）では，AIが教員より低いスコアを付ける傾向があることが分かった。なお，評価の信頼性（安定性）を確認する目的で教員，AIそれぞれの内部評価の安定性を算出したICC（Intraclass Correlation

Coefficient)においては, AI評価では形式的観点では高いICCが得られた一方, AttitudeではICC = .200と低い値を示した。教員評価においても観点によりICCの差が見られたが, 概ね中〜高の安定性を示している。このことは, AIが学習者の内面の意図や学習姿勢などを評価することが困難であることを示しており, Xiao and Zhi(2023)や Polakova and Ivenz(2024)の報告と一致する。

RQ2:評価観点ごとに, 生成 AI に特有の過小または過大評価の傾向は見られるか。
➡AIは一部の観点で一貫して教員より厳しい傾向を示した。

箱ひげ図による視覚化(図1, 図2)では, 英検準拠の4観点(内容, 構成, 語彙, 文法)すべてにおいて, AIの方が教員評価よりもやや低いスコアを付ける傾向が確認された。特に文法においては, 差分スコアの中央値が大きく教員評価優位に偏っており, AIの評価基準がより厳格である可能性がある。また, 新学習指導要領に準拠した観点でも, 知識, 態度においてAIは一貫して過小評価する傾向を示しており, 特に主観的評価におけるAIの限界を浮き彫りにした。さらに, 両評価の関連性をピアソンの積率相関係数によって検証したところ, 英検準拠ルーブリックでは内容(r = .790, p < .001), 構成(r = .756, p < .001), 語彙(r = .716, p < .001), 文法(r = .812, p < .001)と, いずれも高い相関が得られた。新学習指導要領準拠ルーブリックにおいても, 知識(r = .739, p < .001), 表現(r = .747, p < .001), 態度(r = .784, p < .001)と, 中程度から高い正の相関が確認された。以上から, AI評価は観点ごとに教員評価と一定の一貫性を有する一方で, 特定の観点においては厳格な評価傾向を示すことが明らかになった。

RQ3:RQ1と2で明らかになった一致度や傾向差を前提として, 高等学校英語教員は生成AIによる英作文評価をどのように受容しているか。
➡「使える場面もあるが, 主観的評価は教員が担うべき」という条件つきの受容が多く見られた。

TAMに基づくアンケート分析では,「知覚された有用性(PU)」と「使いやすさ(PEOU)」はいずれも平均3.5以上と高く評価され, AIツールとしてのポテンシャルは一定程度認識されていることが分かった。一方,「行動意図(BI)」は3.3とやや低めであり, 回帰分析ではPUがBIに統計的に有意な負の影響(β = −0.651, p = .022)を与える結果となった。これは, AIを「理論上は有用だと考えているが, 実際の授業では積極的に使いたいとは限らない」というジレンマを反映している可能性がある。自由記述では,「AI評価は補助的に使用するのが現実的であり, 主観的観点の判断や生徒理解には教員の介入が不可欠」とする声が多く, AIの過信を戒める姿勢が顕著だった。また,「プロンプトによって出力結果が変わるため, 教員自身にプロンプト設計の知識が必要」といった指摘もあり, AI活用における教員の新たなリテラシーの重要性も示唆された。

以上の結果から, 生成AIによる英作文評価は, 形式的観点においては信頼性と一貫性を確保しうるものの, 主観的, 意味的観点においては限界があり, 全面的な代替は困難であることが分かった。したがって, AIはあくまで補助的な評価ツールとして活用し, 評価の中心は引き続き教員が担う必要がある。また, AIの評価結果を学習者自身が振り返る材料として活用することで, 深い学びを促進する可能性もある。今後は,「AI評価をどう授業設計に組み込むか」「プロンプト設計力をどう育成するか」が重要な課題となるだろう。

5 結論および今後の課題

本研究では, 高等学校英語教育におけるAI活用の可能性と課題を明らかにすることを目的に, 生成AIによる英作文評価に対して, 教員との観点別比較, ICC分析, TAMアンケートを通じた意識調査を実施した。分析の結果, AIは文法, 語彙, 構成といった形式的観点において非常に高い評価の一貫性を示し, 教育現場においてこれらの観点の一部をAIが代替する可能性が示唆された。一方で, 意味内容や態度といった文脈的, 主観的判断を要する観点では, 評価のばらつきや限界が確認され, AIによる単独評価には慎重

な対応が求められることも明らかになった。教員アンケートの結果からは，AI評価の有用性や効率性に対する期待と同時に，評価の正確性や教育的責任に対する慎重な姿勢がうかがえた。AIの導入が教員の業務負担を軽減する一方で，学習者理解や指導の質の担保は人間が担うべきだという認識が多くの教員に共有されていた。特に，プロンプト設計やAIの出力の取扱いに関する教員の理解とスキルが，AI評価の信頼性と教育効果を大きく左右する可能性が示されたことは重要である。

今後の課題としては，第一に，複数のAIモデルを用いた比較研究を通じて，モデルごとの評価傾向や性能の違いを明らかにすること。第二に，AIの効果的な利活用に必要なプロンプト設計技術の向上を図り，教員が適切にAIを活用できる体制を構築すること。第三に，AIによる評価が学習者の表現力や学習意欲，思考の深化といった側面に中長期的に与える影響を検証する実証研究を継続的に行うことが挙げられる。また，本研究では英検準拠および新学習指導要領準拠の二種類のルーブリックを作成し，それらを基にAIと教員による評価を行ったが，参加者が評価に迷う場面が少なからず存在した。Stanley（2021）が「ルーブリックに関する知識の中で最も重要なことは，そのルーブリックが良いものか悪いものかを見極める能力である」と指摘するように，評価の目的に応じて適切かつ妥当なルーブリックを設計できる技術が，今後のパフォーマンス評価に関連した生成AI評価研究や教育実践において不可欠である。

謝辞

本研究を進めるにあたり，関西大学の竹内理教授から多大なるご指導とご助言を賜りました。研究活動に行き詰まっていた私に対し，的確なご指摘と温かな励ましを繰り返しくださり，再び研究に向き合う力を与えてくださいました。研究の構想から分析手法に至るまで，先生のご助言を通じて多くを学び，本研究を進めるうえで大きな支えとなりました。心から感謝申し上げます。あわせて，本研究に英作文を提供してくださった室戸高校の生徒の皆さん，ならびにご多忙の中にもかかわらず，評価およびアンケート調査に取り組んでくださった英語教員の先生方にも，深く感謝申し上げます。皆様のご協力と率直なご意見は，本研究にとってかけがえのない支えとなり，貴重な示唆を与えてくださいました。さらに，本研究は公益財団法人 日本英語検定協会による「英検研究助成制度」の助成を受けて実施されたことをここに記し，厚く御礼申し上げます。

倫理的配慮に関る指針

本研究は，高等学校英語教員を対象にAIによる英作文評価の有効性を調査するものであり，研究目的，方法，個人情報の取扱い等を記載した説明文書を事前に配布し，文書による同意を得て実施した。評価対象の英作文は匿名化し，個人が特定されないよう配慮した。AI評価にはChatGPTを用い，出力内容は教育的目的に限って使用した。教員アンケートも無記名で実施し，統計的に処理・分析した。

引用文献

Asli, N. F., Mohd Matore, M. E. E., & Md Yunus, M. (2024). Construct validity of primary trait writing rubrics based on assessment use argument (AUA) validation framework. *Heliyon, 10*(22), e40053-. https://doi.org/10.1016/j.heliyon.2024.e40053

Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. Ascd.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* Longmans, Green.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Chocarro, R., Cortiñas, M., & Marcos-Matás, G. (2023). Teachers' attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics. *Educational Studies, 49*(2), 295–313. https://doi.org/10.1080/03055698.2020.1850426

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

ElEbyary, K., Shabara, R., & Boraie, D. (2024). The differential role of AI-operated WCF in L2 students' noticing of errors and its impact on writing scores. *Language Testing in Asia, 14*(1), 59–24. https://doi.org/10.1186/s40468-024-00312-1

Ellis, R., Loewen, S., & Erlam, R. (2006). IMPLICIT AND EXPLICIT CORRECTIVE FEEDBACK AND THE ACQUISITION OF L2 GRAMMAR. *Studies in Second Language Acquisition, 28*(2), 339–368. https://doi.org/10.1017/S0272263106060141

Fan, Y., Tan, S., & Lim, G. Y. W. (2024). EAP teacher feedback in the age of AI: Supporting first-year students in EFL disciplinary writing. *Australian Journal of Applied Linguistics (Online), 7*(3), 1943-. https://doi.org/10.29140/ajal.v7n3.1943

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*(3), 161–184. https://doi.org/10.1016/S1060-3743(01)00039-X

Fuller, L. P. , & Bixby, C. (2024). The Theoretical and Practical Implications of OpenAI System Rubric Assessment and Feedback on Higher Education Written Assignments. *American Journal of Educational Research, 12*(4), 147-158.

Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System (Linköping), 126,* 103502-. https://doi.org/10.1016/j.system.2024.103502

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education (Lausanne), 7.* https://doi.org/10.3389/feduc.2022.983055

Kim, Haeun & Baghestani, Shireen & Yin, Shuhui & Karatay, Yasin & Kurt, Sebnem & Beck, Jeanne & Karatay, Leyla. (2024). ChatGPT for Writing Evaluation: Examining the Accuracy and Reliability of AI-Generated Scores Compared to Human Raters. 10.31274/isudp.2024.154.06.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6,* 100225.

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review, 11*(3), 25-37.

Lai, C. Y., Cheung, K. Y., & Chan, C. S. (2023). Exploring the role of intrinsic motivation in ChatGPT adoption to support active learning: An extension of the technology acceptance model. *Computers and Education. Artificial Intelligence, 5,* 100178-. https://doi.org/10.1016/j.caeai.2023.100178

Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology.* https://doi.org/10.14742/ajet.9463

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System (Linköping), 127,* 103529-. https://doi.org/10.1016/j.system.2024.103529

文部科学省 . (2024). *初等中等教育段階における生成AIの利活用に関するガイドライン(Ver.2.0)*. https://www.mext.go.jp/content/20241226-mxt_shuukyo02-000030823_001.pdf

Mohammed, S. J., & Khalid, M. W. (2025). Under the world of AI-generated feedback on writing: mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. *Language Testing in Asia, 15*(1), 7–26. https://doi.org/10.1186/s40468-025-00343-2

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language learning, 64*(4), 878-912.

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1). https://doi.org/10.1080/2331186X.2024.2410101

Pozdniakov, S., Brazil, J., Abdi, S., Bakharia, A., Sadiq, S., Gašević, D., Denny, P., & Khosravi, H. (2024). Large language models meet user interfaces: The case of provisioning feedback. *Computers and Education. Artificial Intelligence, 7*, 100289-. https://doi.org/10.1016/j.caeai.2024.100289

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait Ratings for Automated Essay Grading. *Educational and Psychological Measurement, 62*(1), 5–18. https://doi.org/10.1177/0013164402062001001

引用文献

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies, 29*(18), 24735-24757. https://doi.org/10.1007/s10639-024-12817-6

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Stanley, T. (2021). Using rubrics for performance-based assessment: A practical guide to evaluating student work. Routledge.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education. Artificial Intelligence, 7,* 100255-. https://doi.org/10.1016/j.caeai.2024.100255

Teimouri, Y. (2017). L2 SELVES, EMOTIONS, AND MOTIVATED BEHAVIORS. *Studies in Second Language Acquisition, 39*(4), 681-709. https://doi.org/10.1017/S0272263116000243

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education. Artificial Intelligence, 7,* 100270-. https://doi.org/10.1016/j.caeai.2024.100270

Tran, T. T. T. (2025). Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences. *Education Sciences, 15*(2), 232-. https://doi.org/10.3390/educsci15020232

Troia, G. (2014). Evidence-based practices for writing instruction (Document No. IC-5). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/tools/innovation-configuration/

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education, 12*(1), 20-32. https://doi.org/10.21449/ijate.1517994

Yang, Yang. (2024). The Reliability of using ChatGPT in Rating EFL Writings. *Shanlax International Journal of Education. 12.* 49-59. 10.34293/education.v12i4.7855.

Yao, Y., Zhu, X., Xiao, L., & Lu, Q. (2025). Secondary school English teachers' application of artificial intelligence-guided chatbot in the provision of feedback on student writing: An activity theory perspective. *Journal of Second Language Writing, 67,* 101179-. https://doi.org/10.1016/j.jslw.2025.101179

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150-166. https://doi.org/10.1111/bjet.13494

VanPatten, B., & Williams, J. (2007). *Theories in second language acquisition.* NY: Routledge.

Xiao, Y., & Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. *Languages (Basel), 8*(3), 212-. https://doi.org/10.3390/languages8030212

Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning, 6*(3), 1482-. https://doi.org/10.29140/tltl.v6n3.1482

資料1：教員の受容意識（TAM分析）アンケート調査項目

TAM 事前アンケート（生成 AI に関する意識）

1. PU 生成 AI は，英作文評価を効率的にすると思う。
2. PU 生成 AI を使うことで，英作文評価精度が向上すると思う。
3. PU 生成 AI による評価は，生徒の英作文へのフィードバックの質を高めると思う。
4. PU 生成 AI による評価は，教員による英作文評価よりも公平だと思う。
5. ATT 英作文の評価に生成 AI を使用することは，教育現場において価値があると思う。
6. ATT 生成 AI を使用することに対してよい印象を持っている。
7. ATT 生成 AI を評価に取り入れることは，教育実践を改善する助けになると思う。
8. ATT 生成 AI を評価に使用することに対して抵抗感を感じる。
9. SI 同僚教員は生成 AI を使用することを推奨していると思う。
10. SI 生成 AI を使わないと周囲から取り残されると感じる。
11. SI 業務の質向上のために生成 AI を使うことが望ましいと思う。
12. SI 業務の効率化のために生成 AI を使うことが望ましいと思う。

TAM 事後（生成 AI 使用後の評価）

13. PU 生成 AI は，英作文評価を効率的にすると思う。
14. PU 生成 AI を使うことで，英作文評価精度が向上すると思う。
15. PU 生成 AI による評価は，生徒の英作文へのフィードバックの質を高めると思う。
16. PU 生成 AI による評価は，教員による英作文評価よりも公平だと思う。
17. PEOU 生成 AI の操作方法は簡単だと感じる。
18. PEOU 英作文評価のプロセスに生成 AI を組み込むのは簡単だと思う。
19. PEOU 生成 AI を使い始める際，特別な知識やスキルは必要ないと思う。
20. PEOU 生成 AI が生成した結果を理解し，活用するのは簡単だと思う。
21. PEOU 業務において，生成 AI を日常的に使用することに対して疑問やストレスを感じる。
22. BI 英作文評価において，生成 AI を積極的に使用したいと思う。
23. BI 生成 AI が評価ツールとして無料で利用可能な場合，すぐに試してみたいと思う。
24. BI 他の教員にも生成 AI の使用を勧めたいと思う。
25. ATT 英作文の評価に生成 AI を使用することは，教育現場において価値があると思う。
26. ATT 生成 AI を使用することに対してよい印象を持っている。
27. ATT 生成 AI を評価に取り入れることは，教育実践を改善する助けになると思う。
28. ATT 生成 AI を評価に使用することに対して抵抗感を感じる。
29. SI 同僚教員は生成 AI を使用することを推奨していると思う。
30. SI 生成 AI を使わないと周囲から取り残されると感じる。
31. SI 業務の質向上のために生成 AI を使うことが望ましいと思う。
32. SI 業務の効率化のために生成 AI を使うことが望ましいと思う。

資料2：教員の受容意識（TAM分析）主要構成要素の集計結果

	PU	PEOU	BI	ATT	SI
T1	5	3	1	4	3.75
T2	3.78	4	3.33	3.75	2.88
T3	4	4	4	3.75	3.5
T4	4	4	3.33	3.75	3.88
T5	3.78	3.75	3.67	4	4.13
T6	4.67	4	3.33	4.13	3.75
T7	4	2.5	3.67	3.38	3.13
T8	4.89	4	3.67	4.13	3.63
T9	4	4	3	3.5	2.88
T10	4.11	3.75	3.33	3.5	3.75
T11	3.89	4.5	3.33	3.5	3
T12	4.56	4.5	3.67	4.38	4.25
T13	4	3.75	3.33	4.38	3.88
T14	3.67	3.25	3.33	3.13	2.38
T15	3.67	3	3.67	3.38	3.5
T16	3.89	3.5	3.33	4.13	2.88
T17	4	3.25	4.33	4.38	4.13
T18	3.44	3	2.67	3.13	2.75
T19	3.78	3.5	3.33	4.13	2.75
T20	4.33	4.75	3.67	4.5	3.75