第37回 研究助成

A 研究部門 ● 報告Ⅴ ● 英語能力テストに関する研究

多肢選択式読解問題における本文の語を含む割合を調整した正答選択肢の調査

― 生成 AI を用いた選択肢の作成方法の検討 ―

研究者:若松 千智　千葉県／千葉大学大学院　在籍

《研究助言者:斉田 智里》

概要

本研究の目的は，多肢選択式読解問題の正答選択肢における「本文の語を含む割合」を調整した場合，項目正答率・自信度・弁別力のそれぞれがどのように変化するのか検証することである。また，テスト作成者である教員の負担軽減のために，生成AIを用いた効率的な選択肢の作成方法を検討した。英検2級の読解問題（オリジナル）と，正答選択肢のみを生成AIで再作成した問題（パラフレーズ）の二種類のテストを国立大学・大学院に所属する65名に実施した。その結果，本文の語を含む割合を低くした正答選択肢を用いた場合，自信度・弁別力ともに有意差は見られなかった。一方で，本文中の語を多く含む正答選択肢を用いた場合と比較して，正答選択肢に含まれる本文中の語の割合が低い場合には，項目正答率が有意に低下することが確認された。また，いずれのテストにおいても高い信頼性が示された。以上より，生成AIを用いたパラフレーズは項目の信頼性や弁別力を損なわずに活用可能であり，効率的な問題作成に有用な可能性が示唆された。

1 はじめに

1.1 学習指導要領

グローバル化が進んでいく中，小学校で外国語が教科化されるなど，英語の重要性は高まっている。平成29年，平成30年告示の学習指導要領解説・外国語編では，生徒が英語に触れる機会を充実するとともに，授業を実際のコミュニケーションの場面にするため，授業は英語で行うことを基本とすると定められており，リーディングにおいても英語を英語のまま理解することは，読みの流暢性の観点からも重要である。卯城（2011）では，英文には必ずそれを書くための目的があり，英文を理解することは読み手が書き手のメッセージを受け取ることであると言及されており，一文一文を理解することができても，必ずしも英文全体のメッセージをつかむことはできないと言われている。英文を理解するためには，語彙や文法の知識だけではなく，背景知識を活性化させたり能動的に読んだりすることで文章全体のメッセージを捉えることが求められる。授業を実際のコミュニケーションの場面にするため，英語の授業は英語で行うことを基本としていることからも，英語で内容を理解することが求められている。そのため，実際の定期テストでも英語を用いての内容理解を問うべきだが，記述式では中学生にとって難易度が高くなってしまう懸念がある。

指導と評価のずれとして「知識及び技能」と「思考力，判断力，表現力等」の境界線の問題が今井（2023）で挙げられている。同文献では，「知識及び技能」と「思考力，判断力，表現力等」は区別が難しいことが指摘されているものの，目的・場面・状況をもとに言語活動を設定し，適切に言語使用することが「思考力，

判断力, 表現力等」であり, 正確さが「知識及び技能」と定義されている。評価方法にパフォーマンステストを取り入れることにより, 紙のテストの限界を補うことは重要であるものの, 4技能のうち読解能力を測るためにはやはり紙のテストの充実が求められる。しかし, 紙の定期テストでは, 観点別評価をするために「知識及び技能」を測る問題と「思考力, 判断力, 表現力等」を測る問題を分ける必要があり, 中でも読解問題における「思考力, 判断力, 表現力等」を測る問題の作成が課題である。

1.2 生成AIについて

ChatGPTが2022年にリリースされて以降, 教育の在り方にも影響が出ている。初等中等教育における生成AIの利活用に関するガイドライン ver.2.0（文部科学省, 2024）によると, 児童生徒が学習に生成AIを利活用する際だけでなく, 教員が生成AIを校務で利活用することに関するガイドラインが示されている。まず, 教育は教員と児童生徒の人格的な触れ合いを通じて行われるものであり, 生成AIが社会インフラの一部となる時代においてより重要になる。しかし, 教員が授業準備や文書の叩き台作成を含む校務において生成AIを利活用することで校務の効率化につなげることができると示されている。さらに, 教員が生成AIの利便性や懸念点を知ることは, 児童生徒の学びをより高度化する観点からも重要であると言われている。教員自身が生成AIの仕組みや活用方法について学ぶことで業務の効率化につなげることができるが, 不正確な情報を出力してしまうハルシネーションのおそれもあるため, 最終的には教員が生成AIからの出力を正しく判断する必要がある。

生成AIには, OpenAI社のChatGPT, Google社のGemini, Microsoft社のCopilotをはじめとして, 数多く開発されている。日本リサーチセンターの調査によると, 生成AIの利用割合は2024年6月から2025年6月までに14.7パーセンテージポイント上昇し, 2025年6月には30.3%（n = 1,394）を占めている。生成AIを使用している人が利用している生成AIの種類は, 上記3種類の中でChatGPTの利用率は20.9%と最も高い。ChatGPTの仕組みについて, 大規模コーパスを使用し, 事前に言語使用の深層学習を行っており, 次に来る確率の高い語やフレーズを生成していると水本（2024）で説明がされている。このことから, 生成AIは言語処理が得意であり, 英語教育においても親和性が高いと言える。日本の大学入学共通テストにおいて生成AIを活用した研究としてWatanabe and Saida（2024）が挙げられる。Watanabe and Saida（2024）では, 大学入学共通テストの英語リーディングの問題を生成AIがどの程度得点できるかを調査した。その結果, ChatGPTの平均得点は受験者平均を大きく上回ることが明らかになった。さらに, ChatGPTは設問の不明瞭さを発見するなど項目の改善に役立つ可能性が示唆されたことから, 英語教員の補助ツールとしてテスト検証や教材作成に有効であると結論づけている。また, 生成AIを使用したテスト作成について検討したShin（2023）では, 生成AIを用いて多肢選択式テストを作成し, 実際に使用されている韓国の大規模試験（CSAT）との比較を50名の英語教員が行った。研究協力者である英語教員は, ChatGPTを使用して作成したテストと実際に使用されている韓国の大規模試験において, 文の流れや表現の自然さを同程度と評価したことが示されている。これらの結果は, 生成AIを使用したテスト作成の可能性を示唆している。

1.3 テスト作成について

テストを検討する際に意識する必要がある妥当性・信頼性・波及効果・実用性の4観点について, 根岸（2017）では以下のように定義している。

- 妥当性……テストが測ろうとしている知識や能力を測っているか
- 信頼性……何度測っても同じような結果が出るか
- 波及効果…テストが学習や指導に及ぼす影響はどのようなものか
- 実用性……テストの作成・実施・採点・解釈が容易かどうか

これらの4観点が十分に反映されたテスト作成が重要だが，近年の教育現場では教員不足や教員の多忙化が著しく，妥当性，信頼性，波及効果の3つが保たれた上での「実用性」の向上が求められる。定期テストの読解問題の課題として，根岸（2017）では，定期テストでは教科書本文を読解問題として出題しているが，すでに読んだことのある文章はテストで改めて読まなくても内容がわかってしまうことが挙げられていることから，教科書や授業では扱っていない初見の文章を使用する必要性があると言える。

また，読解力を測る手段としてクローズテスト，多肢選択式テスト，組み合わせ法，空所補充，真偽テスト，要約テストなどがあるが，その中の一つである多肢選択式テストは，作成に時間や手間がかかるものの採点が容易であるため，実用英語技能検定，IELTS，TOEFL，TOEICなどの多くの大規模テストで使用されている。一方，多肢選択式テストのデメリットとして，当て推量の可能性があることが挙げられている。多肢選択式の読解問題と記述式（Open-Ended）の読解問題において，項目の種類による難易度への影響を調査したUshiro et al.（2008）では，多肢選択式テストは記述式テストと比べて，点数が予測しにくいことが明らかになっている。このことから，記述式では理解できていないものを産出することができない反面，選択式のテストは分からなくても当たってしまったり，なんとなく当ててしまったりするおそれがあると言える。そのような「当て推量」を防ぐために有効な手段として，もっともらしい錯乱肢を作成することが挙げられており，テスト受験者の能力によって思わず選んでしまう，引きつけられやすさを指す「魅力度」が変化する選択肢の検討が求められる（若林・根岸，1993）。

多肢選択式読解問題における質問タイプを分析した清水（2008）では，質問タイプを大きく上位レベル処理と下位レベル処理の2つに分けている。上位レベル処理の質問タイプとして，推論質問，テーマ質問，文章構造質問が挙げられており，これらは下位レベル処理の質問タイプであるパラフレーズ質問よりも処理が難しいことが示されている。清水（2008）でのパラフレーズ質問は，文章内で記されている文章を言い換えたものであり，先述した今井（2023）で言われている「知識及び技能」の定義である「言語使用の正確さ」を測る項目だと判断できる。「言語使用の正確さ」に加えて，より上位レベルの処理を要する問題を取り入れることにより，「思考力，判断力，表現力等」の定義として示されていた「言語使用の適切さ」を測ることができるだろう。Hasegawa（2017）では，正誤問題を，テキストに明示されているか，もしくは暗示されているかについて，1：推論が必要か，2：2文以上を参照する必要があるかの2つの基準で分類している。テキストに暗示されている内容に関する質問は，清水（2008）の上位レベル処理の質問タイプであり，「思考力，判断力，表現力等」を測る質問であると言える。

1.4 選択肢について

多肢選択式テストでの選択肢は，正答の選択肢と誤りの選択肢である錯乱肢の2種類で構成される。Ushiro et al.（2007）では，TOEFLのテストを使用し，大学生114名を対象に解答者が錯乱肢を選びやすくなる要素を検討しており，錯乱肢のもっともらしさ（plausibility）について，語数や構文などの定量的なデータで判断することは難しく，問題タイプによって異なる要素が確認されている。また，どのような錯乱肢が受験者の能力を正確に識別できるかについて調査した研究として，要約タイプの問題を使用して日本人学習者が要約の過程で起こしがちな典型的なエラーである3種類の錯乱肢を比較したTerao and Ishii（2020）が挙げられる。パッセージの要点を捉える多肢選択式の問題において，1：重要な情報が欠けていたり，不必要な情報が含まれているdeletion，2：本文中の例をそのまま抜き出したり，一般化しすぎた表現を使用したgeneralization，3：段落全体ではなく，一部の内容しか要約をしていなかったり，書き手の視点とは異なる記述をしているintegration，の3種類の錯乱肢を使用した。その結果，重要な情報が欠けている錯乱肢（1：deletion）と詳細な例をそのまま抜き出した要約（3：integration）はテスト受験者の能力を正確に識別することが明らかになっていた。

また，清水（2008）により分類された質問タイプの一つであるパラフレーズ質問に焦点を当てた研究（政所，2018）では，多肢選択式読解問題のパラフレーズ質問をパラフレージング・レベルを用いて，錯

乱肢だけでなく全ての選択肢において分析を行っている。TOEIC, TOEFL, 英検2級の3つのテストの読解問題を使用し, それぞれのテストの選択肢が本文の語をどの程度使用しているかを調査した上でテストを実施した。テスト結果から, テスト得点群によって正答・誤答選択肢の組み合わせによる選択率と魅力度への影響が異なること, 特にそのテストで測る能力が低い人たち(低位群)は正答選択肢が本文の語をあまり含まない場合に, より本文の語を含む錯乱肢を選ぶ傾向があることが明らかにされている。

これらの研究から, 思わず選んでしまうもっともらしい錯乱肢の要素は問題タイプに依存するものの, 能力によって本文の語を含んでいる割合が選ばれやすさの要因となる可能性がある。そこで, 本研究では表層的な要素である本文の語を含んでいる割合を調整して調査を行うこととした。

1.5 本研究

本研究において使用する言葉の定義について, 項目正答率は正解を1, 不正解を0として算出したテストの得点率を指し, 自信度は, 低中高の3段階で, 低を1, 中を2, 高を3として算出した数字を指す。今回扱うテストで測る能力をどの程度識別することができるかを示す弁別力においては, テスト全体の得点とその項目の得点の相関を表す点双列相関係数の値を弁別力の指標として扱う。

本研究では, 正答選択肢における本文の内容語を言い換えた割合を調整した場合, 項目正答率, テスト受験者の正答選択肢に対する自信度, 弁別力のそれぞれがどのように変化するのかを検証する。また, テスト作成者である教員の負担軽減のために, 多肢選択式テストの課題である「作成に時間がかかること」を克服するツールとして生成AIを用い, 効率的な選択肢の作成方法を検討することを目的とし, テストの充実の一助となるよう, 研究課題を以下に3つ設定した。

本文の語を使用する割合がより低い正答選択肢を含む項目は, 本文の語を使用する割合がより高い正答選択肢を含む項目と比べて,

研究課題 1　項目正答率がどの程度変化するのか明らかにすること。

研究課題 2　受験者の項目に対する自信度がどの程度変化するのか明らかにすること。

研究課題 3　受験者の能力を識別する弁別力はどの程度変化するのか明らかにすること。

これらの研究課題に対して, 表面的な難易度が上がるため, 研究課題1における項目正答率と研究課題2における自信度は下がり, 研究課題3における弁別力は上がると仮説を立てた。

2 方法

2.1 マテリアル

テストは実用英語技能検定(以下, 英検)2級2023年度第3回実施分(大問3:A, B, C), 2024年度第1回実施分(大問3:A, B), 2024年度第2回実施分(大問3:A, B)の計3回分を使用した。英検は2024年度から出題形式が変更となり, ライティング問題が1題追加された影響でリーディングの問題数が減っているため, 2023年度実施分と2024年度実施分の読解問題の数が異なっている。それぞれ大問3の文章を読んだ上でいくつかの質問に答える形式の読解問題のみを抽出し, 7つの文章題で小問は合計で28項目使用した。文章

題7つは通し番号をA～Gとし，テストは2種類使用した。1つ目は英検の問題をそのまま使用したもの（オリジナル），2つ目は英検の文章，問題文，錯乱肢はそのまま使用し，正答の選択肢のみパラフレーズしたもの（パラフレーズ）である。それぞれに対して，解答の自信度を調べるために，一つ一つの設問に対して3段階で自信度を尋ねた。

2.2 知識及び技能を測る項目と思考力，判断力，表現力等を測る項目の分類

知識及び技能を測る項目か思考力，判断力，表現力等を測る項目かという分類の際には，英語教育を専門とする大学院生が参加し，2名で分類を行った。分類基準は，Hasegawa（2017），清水（2005），国立教育政策研究所 教育課程研究センター（2020）を参考にし，(1)代名詞が表すものも含めて正答にたどり着くために2文以上参照する必要がある，(2)推論が必要である，(3)段落や文章全体の主題（テーマ）を問う質問である，(4)必要な情報を読み取り，書き手の意図，概要，要点を捉える問題である，の4つに定めた。上記の4つの基準のうち，どれか一つ以上該当する場合に思考力，判断力，表現力等を測る項目とした。全ての項目を2名で分類したところ，一致率は75%だった。2名で一致しなかった項目は話し合うことにより全ての項目を分類したところ，知識及び技能を測る項目は16項目，思考力，判断力，表現力等を測る項目は12項目だった（表1参照）。

■表1: 知識及び技能と思考力，判断力，表現力等の問題分類

知識及び技能	思考力，判断力，表現力等
16項目	12項目

2.3 選択肢における本文の内容語を使用した割合の調査

政所（2018）を参照し，本文に使用されている内容語が選択肢に占める割合を事前に調査した。内容語には名詞・副詞・形容詞・一般動詞を含め，機能語にはbe動詞・助詞・助動詞・接続詞・冠詞・前置詞・代名詞・疑問詞を含めて分類を行った。一つの単語に複数の品詞がある場合は，文章で使用されている品詞を基に分類を行った。計算方法は，(GL（General Links：本文に3回以上登場する内容語）/ 選択肢の総語数）＋（UL（Unique Links：本文に1，2回登場する内容語）/ 選択肢の総語数）＝パラフレージング・レベルとした（政所，2018を参照）。合計が50%以上の場合はNear Copy（NC），20～49%の場合はMinimal Revision（MIR），1～19%の場合はModerate Revision（MOR），0%の場合はSubstantial Revision（SR）として分類を行った（表2参照）。

■表2: パラフレージング・レベルの分類

パラフレージング・レベル	UL，GLの語数の合計が総語数に占める割合
Near Copy（NC）	50%以上
Minimal Revision（MIR）	20～49%
Moderate Revision（MOR）	1～19%
Substantial Revision（SR）	0%

（注）政所（2018）p. 89表2を参照

2.4 生成AIでの選択肢作成

ChatGPT-4oを使用し，正答選択肢のパラフレーズを行った。使用したプロンプトは「(正答選択肢を入れる) Create three paraphrased sentences of the above sentences using CEFR B1 level vocabulary.」である。本文を入力して正答選択肢が本文の語を含む割合を調整するプロンプトも検討したが，当初は生成AIは数字を正確に扱うことができないと言われていたため，このようなプロンプトを使用した。また，生成AIに馴染みのない教員も手軽に使うことができるよう，プロンプトの指定は簡素にしている。なお，事前に一つの文章(3項目)の正答選択肢について，前述のプロンプトで3つのパラフレーズ案を作成し，本文の内容語を含む割合を比較・検討し，十分にパラフレーズされることが確認できたため，プロンプトでは「3つの選択肢」としている。このプロンプトでは，主語や他の錯乱肢にはない部分を自動的に補ってしまうという課題があったが，それらを削除してそのまま使用できる場合や，文頭にToを補えば使用できる場合はそれらの選択肢を採用した(例：It was noticed that bonobos…とパラフレーズが出力された場合，錯乱肢がbonobosから始まっている場合は，It was noticed that の部分を削除したbonobosから始まる文章の選択肢として採用する)。しかし，文法的に修正が必要な場合，語彙を変えなければならない場合は，その選択肢は採用しなかった(1：主語が三単現であり，助動詞(will)から始まる選択肢をパラフレーズした際に，主語が補われてI am going to…になっている場合は… is going to …のように語彙を変えなければならないもの(項目1)。2：主語が複数形の場合にwas not enough…となっており，were not enough…のように語彙を変えなければならないもの(項目18))。

これらの採用しなかった2つの選択肢以外の全ての選択肢について，パラフレージング・レベルと本文の内容語を使用している割合(%)を計算し，本文の内容語を使用している割合(%)が最も低いものを採用した。選択肢の総語数と内容語の種類(people，English等の名詞)によっては，パラフレーズする語彙には限界があること，また，短い選択肢は本文の内容語を使用した割合が高くなりやすいため，本文の内容語を使用している割合がオリジナルの正答選択肢よりも高い場合があった。その場合には，もう一度出力させると作成者側の手間が増えてしまい，研究目的の一つである「選択肢を効率的に作成すること」を達成できなくなってしまうことから，パラフレーズされた選択肢のうち最も本文の内容語が含まれている割合が低いものを使用した。

■表3: 生成AIを使用して言い換えをした選択肢の一例

	GL	UL	総語数	パラフレージング・レベル
出力1	3	0	9	33.3%
出力2	4	2	10	60.0%
出力3(採用)	3	0	11	27.3%

オリジナルの正答選択肢：Keep an eye on how words are used in the English language.
- 出力1　：Pay attention to how words are used in English.
- 出力2　：Watch carefully how people use words in the English language.
- 出力3(採用)　：Notice how words are used when speaking or writing in English.

参考までに，オリジナルテストの選択肢のパラフレーズの度合いを調べたところ，表4，5の通りになった。合計で28項目使用し，選択肢は各問4つのため，選択肢の合計は112だった。政所(2018)では，選択肢の内容語のうち，本文で使用されている割合を算出していたため，その方法を採用した(2.3 参照)。

■表4: オリジナルテストの全ての選択肢におけるパラフレージング・レベル

	2023年度第3回	2024年度第1回	2024年度第2回	合計	割合
NC	10	6	4	20	17.9%
MIR	29	19	23	71	63.4%
MOR	7	4	4	15	13.4%
SR	2	3	1	6	5.4%
合計	48	32	32	112	100%

■表5: オリジナルテストの正答選択肢におけるパラフレージング・レベル

	2023年度第3回	2024年度第1回	2024年度第2回	合計	割合
NC	3	0	2	5	17.9%
MIR	8	7	5	20	71.4%
MOR	1	0	1	2	7.1%
SR	0	1	0	1	3.6%
合計	12	8	8	28	100%

オリジナルテストの全ての選択肢のパラフレージングレベル（表4参照）と正答選択肢のパラフレージング・レベル（表5参照）は，それぞれMIRが最も多く，次にNCが多いという政所（2018）の結果と一致した。政所（2018）では，清水（2005）で分類されている質問の一つであるパラフレーズ質問のみに焦点を当てている。本研究では質問の種類での分類は行っていないが，全体的に同様のパラフレージング・レベルの傾向があることが確認できた。

また，事前に英語教育を専攻している大学院生2名に正答選択肢をパラフレーズしたテスト問題を解いてもらい，選択肢に不自然な点がないかの点検と解答時間の記録を依頼した。選択肢では，メールを読み取る形式の問題で，To give her your best wishes for…と出力されていた選択肢において，メールの相手ではなく読み手を指すyourが補われていた部分について指摘を受けたため，yourを削除して使用した。解答時間については，依頼した2名は全体を通して30分から40分だったが，英語を専門としていない人も調査の対象とするため余裕を持って1時間で設定した。

2.4.1 調査協力者

協力者は国立大学・大学院に所属する学生65名であり，所属学部は教育学部・工学部・理学部・国際教養学部・園芸学部・法政経学部等，多岐にわたる。

2.4.2 実施方法

2025年1月中旬から2月上旬にかけてテストを実施した。大学関係者のみにアクセスを制限した上でGoogle Formで参加者を募り，協力者の希望時間に合わせて16グループに分けて実施した。テスト実施前に，研究参加は任意であり途中退室も可能であることや，データは統計的に処理し個人は特定できない旨を記した書面を配布した。その書面を実施者が読み上げて説明をし，参加者の協力への同意を署名を以って確認した。今回は研究テーマや研究目的を事前に伝えると結果に影響を及ぼす可能性があるため，事前の同意書の段階では研究テーマや研究目的を伝えていない。また，テストには匿名での参加が可能である

点を伝えたところ，全員が匿名での参加を選んだ。

制限時間は1時間とし，オリジナル（選択肢そのままのもの），パラフレーズ（正答の選択肢のみ生成AIでパラフレーズしたもの）の2種類を使用し，7つの文章題でカウンターバランスをとった（表6参照）。テストの表紙には，全ての問題を解き切ること，解答を書いたら1項目ずつ自信度に丸をつけること，提示された順番通りに問題に解答すること，の3点について守るように依頼する文章を載せており，テスト前にそれらの説明が書かれた表紙を見せながら実施者が読み上げて説明を行った。英文のジャンルが2種類あり，メールの文章が3つと説明文が4つであること，文章題一つにつき7分くらいの目安で行えば1時間以内で終了することを事前に説明した。事前に問題の全ての分量や内容を確認することは可能であるとし，解く際には前から順番に解くことを徹底してもらうよう伝えた。同時に1名から8名でテストを行った。同時に参加した協力者が1〜2名であり，かつ全員が1時間以内で解き終えた場合は制限時間の1時間を待たずに終了した。

■表6: 問題タイプ別受験者数

Type	問題順番	オリジナル	パラフレーズ
1	A→B→C→D→E→F→G	4	5
2	B→C→D→E→F→G→A	5	5
3	C→D→E→F→G→A→B	5	5
4	D→E→F→G→A→B→C	5	5
5	E→F→G→A→B→C→D	5	4
6	F→G→A→B→C→D→E	4	4
7	G→A→B→C→D→E→F	4	4
		32名（注1）	32名

（注1）時間内に解答ができなかった1名を除いた数

2.4.3 分析方法

オリジナル（選択肢そのままのもの）を解いた人数は33名，パラフレーズ（正答の選択肢のみ生成AIでパラフレーズしたもの）を解いた人数は32名だった。このうち，オリジナルの問題を解いた1名は制限時間内での解答ができなかったため，分析から除外した。また，パラフレーズを解いた人のうち，1名が1項目について解答なし，1名が1項目について解答はしているものの自信度の解答がなかった。前者は問題用紙に下線が引いてあり，読んだ形跡があった。この場合，実際のテストでもわからない問題はとばすという方略があるため，質問を誤って飛ばしたのではなく意図的に飛ばしたとみなし誤りとして扱った。分析は，正答率と自信度のそれぞれにおいて対応のないt検定を行った。なお，今回は正答率と自信度の関係については分析対象としていない。

3 結果と考察

R（R Core Team, 2025）を使用して信頼性係数を出したところ，オリジナルは$\alpha = .85$，パラフレーズは$\alpha = .87$であり，どちらも高い信頼性が確認できた。

3.1 パラフレーズした選択肢

オリジナルとパラフレーズのそれぞれの正答選択肢における本文の語を含む割合を一覧にして以下に示す。

■表7: 2023年度第3回実施分

項目番号	A			B				C				
	1	2	3	4	5	6	7	8	9	10	11	12
オリジナル	36.4%	10.0%	50.0%	33.3%	23.1%	22.2%	30.8%	20.0%	55.6%	23.1%	54.5%	42.9%
パラフレーズ	22.2%	7.1%	33.3%	27.3%	25.0%	0%	20.0%	12.5%	30.0%	18.2%	41.7%	35.7%

■表8: 2024年度第1回実施分

項目番号	D			E				
	13	14	15	16	17	18	19	20
オリジナル	37.5%	0%	22.2%	30.8%	28.6%	33.3%	37.5%	38.5%
パラフレーズ	14.3%	0%	16.7%	18.2%	18.2%	25.0%	0%	28.6%

■表9: 2024年度第2回実施分

項目番号	F			G				
	21	22	23	24	25	26	27	28
オリジナル	35.7%	30.0%	14.3%	30.8%	40.0%	66.7%	30.8%	60.0%
パラフレーズ	30.8%	18.2%	12.5%	18.2%	33.3%	27.3%	8.3%	33.3%

■表10: 本文の語を使用した割合ごとの選択肢分布

割合	オリジナル		パラフレーズ	
	項目番号	項目数	項目番号	項目数
100%		0		0
99%−90%		0		0
89%−80%		0		0
79%−70%		0		0
69%−60%	26,28	2		0
59%−50%	3,9,11	3		0
49%−40%	12,25	2	11	1
39%−30%	1,4,7,13,16,18,19,20,21,22,24,27	11	3,9,12,21,25,28	7
29%−20%	5,6,8,10,15,17	7	1,4,5,7,18,20,26	6
19%−10%	2,23	2	8,10,13,15,16,17,22,23,24	9
9−1%		0	2,27	2
0%	14	1	6,14,19	3
		28		28

表7〜表10のデータから，項目5, 項目14の合計2つを除いてパラフレーズの方が正答選択肢における本文の語を含む割合が低くなっていることが明らかになった。項目14についてはオリジナルの正答選択肢が本文の語を含む割合が0%だったため，それ以上に本文の語を含む割合を下げることができなかった。項目5についてはオリジナルとパラフレーズのそれぞれの正答選択肢に対して，本文の語はそれぞれ3語ずつ含まれており，オリジナルの選択肢の総語数が13語に対し，パラフレーズの総語数が12語と少なかったことが，パラフレーズの正答選択肢における本文の語を含む割合が高くなった原因である。項目5の正

答選択肢を以下に示す。

オリジナル　:a large number of people are using the word in the same way.(13語)

パラフレーズ:a lot of people are using the word in a similar way.(12語)

オリジナルではpeople, word, sameの3語が本文の語を使用しており, パラフレーズではlot, people, wordの3語が本文の語を使用している。

3.2 テスト受験協力者の外部英語試験のスコア

2種類のテストは正答選択肢のみ異なるが問題文が同じであり, 同じ人が2種類のテストを解くことはできないため, オリジナル, パラフレーズを解いたそれぞれの集団が同質となるように最新の英語の資格試験のスコアを事前に聞いた上で割り振った(表11参照)。

英検以外の外部試験の場合, 4技能を測っているもの(IELTS, TOEFL等)であればCEFR対応表を参照して対応する英検の級とみなした。また, 例外として不合格のスコアが書かれていた場合, その記載されたスコアに相当する級として扱った(例:英検準1級不合格スコア2100点は英検2級相当として扱う)。

TOEICはリーディング, リスニングの2技能のみのスコアだったため, 英検の取得級やその他の4技能を測るテストのスコアとは別に, 2つのグループで均等になるように分けた(表12参照)。

■表11: 外部英語試験のスコア別の人数

	英検1級相当	英検準1級相当	英検2級相当	英検準2級相当	TOEIC	取得歴なし	計
オリジナル	0	13	7	0	6	6	32
パラフレーズ	1	12	8	1	5	5	32

■表12: TOEICスコアの得点分布(注:外部試験の受験経験がTOEICのみの協力者)

点数	500～	600～	700～	800～	合計
オリジナル	2	1	2	1	6名
パラフレーズ	2	2	0	1	5名

3.3 得点

問題は合計28項目使用し, 正答の選択肢を選ぶことができていた場合1点, 誤りの選択肢を選んだ場合は0点とし, 全部で28点満点として処理した。表13のテストの得点の記述統計を参照すると, どちらも天井効果が見られた($M + SD$:オリジナル28.82, パラフレーズ28.20)。テスト受験協力者が最近取得した英語資格のアンケートからも, 全体的に英語への関心や意識が高いことが影響していると考えられる。

■表13: テストの得点

	N	*M*	*SD*	*Max*	*Min*
オリジナル	32	25.56	3.25	28	16
パラフレーズ	32	23.78	4.41	28	11

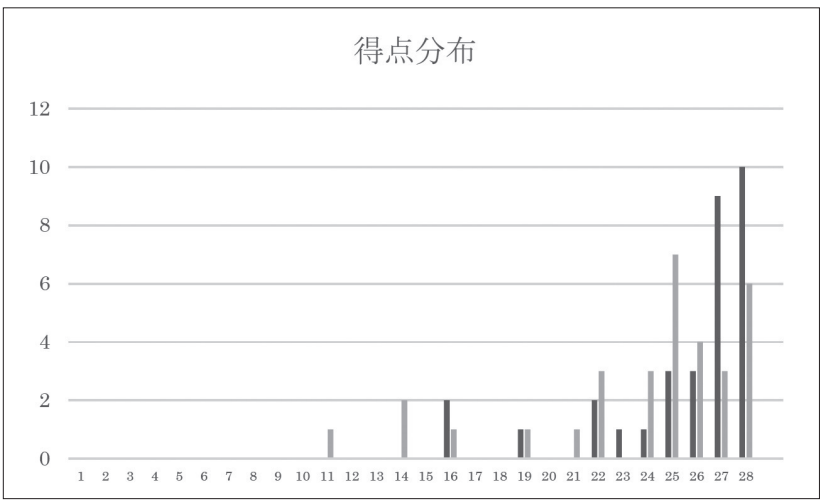
(注)満点は28点

オリジナルとパラフレーズの受験者の得点において対応のないt検定を行ったところ, 有意差がなかった(t = 1.81, df = 62, p = .076)。一方, 項目正答率についても同様に対応のないt検定を行ったところ, 有意差があった(t = 2.56, df = 54, p = .013)。表24に記載した項目正答率の平均では, オリジナルでは91.3%

に対して，パラフレーズでは84.9%であり，全体的にパラフレーズの項目の難易度が高いことが示された。

■表14: 得点分布

	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	合計
オリジナル	0	0	0	0	0	2	0	0	1	0	0	2	1	1	3	3	9	10	32
パラフレーズ	1	0	0	2	0	1	0	0	1	0	1	3	0	3	7	4	3	6	32



■図1：得点分布
(注)濃:オリジナル，薄:パラフレーズ

全28項目のうち，オリジナルを解いた受験者全員が正解だった項目が，項目6，項目9，項目10，項目14，項目24，項目26，項目28の合計7項目だった。パラフレーズを解いた受験者全員が正解だった項目は，項目9，項目24，項目25の合計3項目だった。オリジナルとパラフレーズの両方で全員が正解した項目は項目9，項目24であり，項目25のみがパラフレーズのみで受験者全員が正解だった。項目25はオリジナルのテストでも誤りを選んだ人は一人のみだったため，項目25においてはオリジナルとパラフレーズの両者の項目において，大きな差はないとみなすことができる。

オリジナルとパラフレーズの項目正答率について対応のない*t*検定をしたところ，*p*値が.05未満の項目は合計2つだった(表15参照)。その2つにおいて，オリジナルの方がパラフレーズよりも項目正答率が有意に高い結果がでた。全体としてパラフレーズの方がオリジナルよりもやや難しい傾向があると言えることから，正答選択肢に含まれる本文の語の割合がより低い場合については難易度がやや上がることがわかる。

■表15: 項目正答率に有意差がある項目

項目番号	項目正答率		2種類のテストの正答率の差(%) (オリジナルーパラフレーズ)	*p*値
	オリジナル	パラフレーズ		
1	93.8	75.0	18.8	.039
22	84.4	62.3	22.1	.049

全体的に受験者の正答率が高く天井効果があり，オリジナルとパラフレーズで正答選択肢の選択率にも大きな差はなかったため，今回は各選択肢の選択率については分析対象外としている。オリジナルとパラフレーズで，異なる錯乱肢を選んだ人数が5人以上異なる項目を探したところ，項目12が該当したため，項目12における正答選択率と錯乱肢との関係については表16に示す。

■表16: 項目12 選択肢別の解答詳細

選択肢	オリジナル			パラフレーズ		
	本文の語を含む割合	選択した受験者数	割合	本文の語を含む割合	選択した受験者数	割合
1	42.9	3	9.4	-	1	3.1
2（正答）	42.9	25	78.1	35.7	23	71.9
3	35.7	4	12.5	-	2	6.3
4	46.2	0	0	-	6	18.8

（注）- はオリジナルと同じ値を表す

項目12
問題文：Which of the following statements is true?
錯乱肢：1. Marie Curie had a lot of difficulty remembering facts when she was a child.
2.（正答選択肢）
3. Marie Curie was taught science by a female professor at a university in Paris.
4. Marie Curie wanted to study science after finding out how X-ray machines worked.
オリジナルの正答選択肢
3. Marie Curie and her sister agreed to help each other to study at university.
パラフレーズの正答選択肢
3. Marie Curie and her sister promised to assist each other in their university education.

項目12の正答選択肢は2であり，オリジナルで正答選択肢を選んだ割合は78.1%，パラフレーズで正答選択肢を選んだ割合は71.9%だった。錯乱肢4は，オリジナルでは選択した人が0名だったのに対しパラフレーズでは6名いた。オリジナルの正答選択肢の本文の語を含む割合が42.9%であり，パラフレーズの正答選択肢の本文の語を含む割合は35.7%だったのに対して，錯乱肢4の本文の語を含む割合は46.2%だった。このことから，オリジナルの正答選択肢と錯乱肢4の本文の語を含む割合には大きな差はなかったが，パラフレーズの正答選択肢よりも錯乱肢4の本文の語を含む割合には10.5パーセンテージポイントの差があり，本文の語を含む割合が高い錯乱肢の方を選んでしまった可能性がある。

■表17: 知識及び技能を測る項目と思考力・判断力・表現力等を測る項目の平均正答率

	オリジナル		パラフレーズ	
	知識及び技能（16項目）	思考力・判断力・表現力等（12項目）	知識及び技能（16項目）	思考力・判断力・表現力等（12項目）
平均項目正答率	93.4	88.5	84.8	85.2

知識及び技能を測る項目と思考力・判断力・表現力等を測る項目について，対応のないt検定を行った結果，オリジナルとパラフレーズともに有意差はなかった（オリジナル：t = 1.5587, df = 26, p = .131，パラフレーズ：t = -0.098693, df = 26, p = 0.922）。項目正答率においては両者に違いは確認できなかった。

3.4 自信度

自信度は項目ごとに低中高の3段階で丸をつける方法を採用した。協力者には3段階の自信度の評価を28項目全て解答してもらい，高を3，中を2，低を1として計算を行ったため，一人当たりの最大値は84になる。表18には，オリジナルとパラフレーズのそれぞれの自信度における記述統計を示す。オリジナルとパラフレーズにおける全ての項目の自信度を対応のない*t*検定をしたところ，有意差がなかった（$t = .81$, $df = 60$, $p = .42$）。問題ごとに有意差を確認したところ，表19の通りになった。

■表18: 2種のテストの自信度

	N	*M*	*SD*	*Max*	*Min*
オリジナル	32	69.13	11.65	83	37
パラフレーズ	30	66.53	13.24	83	32

（注）全28問を自信度3段階で解答し，一人当たりの最大値は84

■表19: 自信度に有意差がある項目

項目番号	自信度 平均		2種類のテストの差（オリジナルーパラフレーズ）	*p*値
	オリジナル	パラフレーズ		
1	2.0	2.4	-0.4	.035
2	2.3	1.8	0.5	.013
6	2.8	2.3	0.5	.0038
13	2.6	2.2	0.4	.045
17	2.1	2.5	-0.4	.023
26	2.7	2.3	0.4	.025

（注）自信度の平均は高=3，中=2，低=1として算出

テスト全体として自信度に有意な差は確認できなかったが，項目ごとの自信度について*t*検定を行った結果，有意差の確認ができた項目は合計6つだった。項目17は，オリジナルとパラフレーズの正答率が同じであり87.5%だったものの（項目正答率については表24参照），自信度の平均はパラフレーズの方が高かった。項目1と項目17を除き，本文の語を含む割合が低い正答選択肢を使用したパラフレーズの問題の方が自信度が低く，その結果に有意差が確認できた。項目1と項目17の具体的な項目と選択肢を以下に示す。自信度について有意差がある項目6項目について，項目1,13の2つは「知識及び技能」を測る項目であり，項目2,6,17,26の4つは「思考力，判断力，表現力等」を測る問題だった。全体の項目の傾向として「知識及び技能」を測る項目が16項目に対して「思考力，判断力，表現力等」を測る項目が12項目であり，「思考力，判断力，表現力等」を測る項目の方が全体に占める数が少ない。しかし，自信度について有意差のある項目6つのうち「思考力，判断力，表現力等」を測る項目は4つあり，「知識及び技能」を測る項目よりも，6つのうちに占める割合が多い。このことから，「思考力，判断力，表現力等」を測る項目に対する自信度は相対的に下がる可能性があると言える。

項目1　知識及び技能を測る問題

問題文：Some employees from the IT department

錯乱肢：1. have been selected to work at several job fairs at different colleges.

2. were asked to give a guest lecture to computer engineering students.

3.（正答選択肢）

4. are being sent to another department to help train new workers.

オリジナルの正答選択肢

3. will take part in an event at a university in Petersburg.

パラフレーズの正答選択肢

3. will join an event at a university in Petersburg.

項目1の正答選択肢は，動詞のみ異なる（オリジナル：take part in，パラフレーズ：join）。オリジナルの方が自信度が有意に低い点については，オリジナルを解いた人のうち誤りの選択肢である2を選んだ2名が自信度を低で選択している。また，正答選択肢を選んだ30名中10名が自信度を低で選択し，加えて9名が自信度を中で選択している（表20, 21を参照）。オリジナルの方が自信度が低い理由として，オリジナルで使用されていたtake part inはコロケーションであったことが挙げられる。加えて，パラフレーズで使用されていたjoinの方がオリジナルで使用されていたtake part inよりも，受験者にとって馴染みのある語彙であったことが影響していると考えられる。

■表20: 項目1の選択肢を選んだ人数

選択肢	選択肢を選んだ人数	
	オリジナル	パラフレーズ
1	0	3
2	2	4
3(正答)	30	24
4	0	1
合計	32	32

■表21: 項目1の自信度の人数分布

自信度	選んだ人数	
	オリジナル	パラフレーズ
低	12	5
中	9	9
高	11	18
合計	32	32

項目17　思考力, 判断力, 表現力等を測る問題
問題文：What is one thing the members of the Arts and Crafts movement believed?
錯乱肢：1. Machines would never be able to make objects as quickly as artists could.
2. Paintings and sculptures were not worth making because they were not useful.
3.（正答選択肢）
4. Designers should specialize in producing just one or two types of items.
オリジナルの正答選択肢
3. The lives of people would improve if they returned to making things by hand.
パラフレーズの正答選択肢
3. If people returned to handmade production, their lives would become better.

項目17は, 本文を2文以上参照した上で, 推論する必要である項目であるため,「思考力, 判断力, 表現力等」を測る項目に分類した（分類方法は2.2を参照）。総語数は, オリジナルが14語でありパラフレーズが11語だった。オリジナルは結論を先に示したのちに条件（if）を示しているのに対し, パラフレーズは条件（if）を示したのちに結論を示しており, パラフレーズの方が日本語の文章の順番に近い形であることが, 自信度が高くなった要因として予想できる。Oxford Advanced Learner's Dictionary第10版によると, オリジナルの正答選択肢に含まれているby handには, (1)機械ではなく人の手によるもの, (2)手書きや手渡しであること, の2つの意味が記載されている。このように, by hand には複数の意味があることがオリジナルの正答選択肢への自信度を下げている要因であると考えられる。また, by hand をWeb Para Newsというコーパス検索サイトで調べると, byのヒットは500件（以上）, handのヒットは178件ある中で, by handのコロケーションのヒットは1件のみだった。別のコーパスサイトとして, 小学校・中学校・高等学校の教科書における語彙を検索することができるTeachLex（Sato, 2025）を使用し, handとby両方の登場頻度とコロケーションを調べた。handの登場頻度は, 全ての教科書を合計して小学校で39回, 中学校で27回であり, byの登場頻度は, 小学校で107回, 中学校で178回だった。語彙同士の結びつきを表しているワードクラウドでも, byとhandの強い結びつきは小学校・中学校・高等学校のいずれの学校種においても確認することができなかった。以上のことから, by handのコロケーションの使用頻度が低く, 自信度を下げている要因として考えられる。一方で, パラフレーズの正答選択肢に使用されているhandmadeという語は日本語でもよく使用されていることから, パラフレーズの方が本文の語を含む割合が低いにもかかわらず, 自信度が高くなったと考えられる。

■表22: 項目17の選択肢を選んだ人数

選択肢	選択肢を選んだ人数	
	オリジナル	パラフレーズ
1	2	2
2	1	0
3（正答）	28	28
4	1	2
合計	32	32

■表23: 項目17の自信度の人数分布

自信度	選んだ人数	
	オリジナル	パラフレーズ
低	9	3
中	11	9
高	12	20
合計	32	32

3.5 弁別力

テスト全体の得点と，その項目の正答率の相関を表す点双列相関係数を求めた（表24を参照）。竹内・水本（2023）を参照し，項目―全体得点相関係数である点双列相関係数は0.30以上であれば能力の弁別ができている項目として結果を解釈した。

オリジナルは7項目について全員が正解，パラフレーズは3項目について全員正解の項目があったため，それらは分析の過程で除外されているが，それ以外の項目の弁別力は表24の通りである。項目1，項目11，項目13，項目17，項目18，項目20，項目23の計7項目については，オリジナルよりもパラフレーズの方が.10以上高い値が出ており，項目2，項目4，項目16，項目22，項目27の計5項目については，パラフレーズよりもオリジナルの方が.10以上高い値が出ている。どちらのテストも概ね弁別力は高く，弁別力が.30未満の項目は，オリジナルでは項目11，項目13，項目15，項目18の計4項目，パラフレーズでは項目4，項目6，項目8，項目16の計4項目だった。また，オリジナルの各項目における弁別力の平均値は，.50に対して，パラフレーズの各項目における弁別力の平均値は .49であり，差はほとんどなかった。この結果から，オリジナルとパラフレーズで，弁別力に大きな差は見られず，弁別力はどちらも同等であることが明らかになった。

■表24: 項目別正答率・弁別力・自信度

項目	項目別正答率		弁別力		自信度	
	オリジナル	パラフレーズ	オリジナル	パラフレーズ	オリジナル	パラフレーズ
1	93.8	75.0	.52	.67	2.0	2.4
2	71.9	65.6	.51	.41	2.3	1.8
3	90.6	84.4	.55	.54	2.5	2.3
4	90.6	87.5	.85	.25	2.6	2.2
5	96.9	87.5	.53	.62	2.7	2.7
6	100.0	87.5	-	.26	2.8	2.3
7	81.3	87.5	.65	.56	2.4	2.5
8	87.5	93.8	.67	.075	2.8	2.7
9	100.0	100.0	-	-	2.9	2.8
10	100.0	90.6	-	.37	2.7	2.6
11	96.9	90.6	.20	.62	2.7	2.5
12	78.1	71.9	.53	.55	2.1	1.9
13	93.8	78.1	.24	.54	2.6	2.2
14	100.0	93.8	-	.57	2.4	2.6
15	96.9	96.9	.031	.40	2.6	2.5
16	87.5	90.6	.47	.28	2.5	2.6
17	87.5	87.5	.53	.69	2.1	2.5
18	87.5	71.9	.27	.40	2.3	1.9
19	84.4	78.1	.55	.44	2.2	2.0
20	90.6	81.3	.42	.77	2.4	2.4
21	93.8	78.1	.48	.47	2.4	2.1
22	84.4	62.5	.87	.52	2.0	2.3
23	71.9	75.0	.36	.56	2.5	2.5
24	100.0	100.0	-	-	2.9	2.9
25	96.9	100.0	.53	-	2.4	2.4
26	100.0	90.6	-	.52	2.7	2.2
27	93.8	93.8	.76	.66	2.4	2.4
28	100.0	84.4	-	.56	2.4	2.2
平均	91.3	84.9	.50	.49	2.5	2.4
SD	8.16	9.99	.20	.16	0.26	0.27

（注）弁別力における - は受験者全員が正解だったため，除外された問題

4 結論

本研究は，正答選択肢における本文の内容語を言い換えた割合を調整した場合，項目正答率・テスト受験者の正答選択肢に対する自信度・弁別力のそれぞれがどのように変化するのかを検証することを目的として行った。また，教育現場での読解問題の作成において，客観的な採点が可能であり，実用性の面で採点が容易であるという特徴を持つ多肢選択式テストに焦点を当て，生成 AI を活用した選択肢の作成方法を検討した。ChatGPT-4o を使用して本文の語を含む割合を調整した正答選択肢を作成し，2種類のテストを合計64名の大学生・大学院生を対象に実施した。2種類のテストはともに信頼性は高く，本文の内容語を含む割合を低くした正答選択肢を使用しても，オリジナルの項目と同等の高い信頼性が得られることが明らかになった。本文の語を使用する割合がより低い正答選択肢を含む項目と，本文の語を使用する割合がより高い正答選択肢を含む項目とを比較した結果について，先に設定した3つの研究課題の結論を以下に示す。

研究課題1の「項目正答率がどの程度変化するのか明らかにすること」について，本文の語を使用する割合を低くした正答選択肢を使用した場合，項目正答率が有意に低くなり，全体的に難易度がやや高くなるという結果が明らかになった。このことから，項目正答率は下がるという仮説は支持された。政所（2018）では，特に低位群では，正答選択肢が本文の語をあまり含まない場合，より本文の語を含む錯乱肢を選ぶ傾向があることが明らかにされている。本研究では天井効果が見られたため能力別の比較はできなかったものの，本文の語を使用する割合が選択肢の選ばれやすさに影響を及ぼす可能性があると言える。

研究課題2の「受験者の項目に対する自信度がどの程度変化するのか明らかにすること」について，自信度は下がるという仮説は部分的に支持された。本文の語を含む割合が低い正答選択肢を使用したパラフレーズの方が自信度は低く，有意差がある項目は28項目中6項目だった。自信度についても，本文の語を含む割合が低い場合には，表面的に難しく見える可能性があり，自信度が下がるという仮説は部分的に支持できると言える。

研究課題3の「受験者の能力を識別する弁別力はどの程度変化するのか明らかにすること」について，弁別力においてはほぼ同等であることが示唆された。そのため，弁別力は上がるという仮説は支持されなかった。

本研究では天井効果が見られたが，オリジナルテストの選択肢は信頼性・項目弁別力ともに十分であり，正答選択肢をパラフレーズしてもオリジナルテストと同等の信頼性や弁別力が保たれることが示唆された。生成 AI を使用してパラフレーズをしても語彙の難易度が著しく上がることもなく，錯乱肢と比較しても違和感のない選択肢を作成することができたと言える。一方で，本研究の結果からは，傾向を確認するための量的なデータを十分に提示することができず，選ばれやすさの要素が問題によって異なる場合があることも示された。そのため，どのような問題で本文の語を含む割合が選ばれやすさの要素になるかについて，さらに研究を行う必要がある。テスト作成で生成 AI を用いて言い換えをする作成方法については，テストに天井効果が見られただけでなく，本研究の結果のみでは一般化するために十分な量的なデータを提示することはできなかったことから，一部支持できると言える。生成 AI を積極的に使用することで，実用性の観点から作成時間も短縮することができる。教員による最終チェックは必要だが，業務の効率化につなげることができるだろう。

最後に，本研究の限界点として，天井効果が見られた点と，語彙レベルの調整の2点が挙げられる。使用した2種類のテストでは受験者の得点が高く，どちらも天井効果が見られたため能力別の傾向を明らかにすることができなかった。語彙レベルについては，英検2級の問題では言い換えができる語彙の量が十分だったため，生成 AI を使用して言い換えた正答選択肢の語彙が著しく難しくなることはなかった。しかし，言い換えをする際に使用できる語彙の量が少ない初級レベルの学習者向けのテストでは，本研究とは異なる結果が出ることが予想できる。今後は生成 AI を用いて作成したテストの妥当性と信頼性について，

読解の本文や錯乱肢の言い換え，異なる語彙レベルでの言い換えをしたテストを用いて検証したい。

謝辞

本研究の実施の機会を与えてくださった公益財団法人 日本英語検定協会の皆さま，選考委員の先生方に厚く御礼申し上げます。特に，本研究を担当してくださいました斉田智里先生には貴重なご助言を賜りましたことを深く感謝申し上げます。また，千葉大学の星野由子先生には，本研究に関して計画・実施・報告書の作成に至るまで親身にご指導いただきましたこと，心から感謝申し上げます。最後に，本調査を実施するにあたり，問題の分類やテストの事前チェックにご協力をいただいた皆さま，並びに調査にご協力をいただいた皆さまに厚く御礼申し上げます。

なお，本報告書に関して，開示すべき利益相反関連事項はありません。

引用文献

Anthony, L & Chujo, K. Web Para News. https://www.antlabsolutions.com/webparanews/ (2025年7月25日閲覧)

Hasegawa, Y. (2017). Analyzing Explicit and Implicit Reading Questions in a Term-Exam: A Case Study. *JLTA Journal, 20,* 57-75.

今井裕之. (2023). 外国語科における三観点評価の課題. *KELES* ジャーナル, 8, 82-86.

国立教育政策研究所教育課程研究センター(2020).『「指導と評価の一体化」のための学習評価に関する参考資料 高等学校外国語』国立教育政策研究所

国立教育政策研究所教育課程研究センター(2020).『「指導と評価の一体化」のための学習評価に関する参考資料 中学校外国語』国立教育政策研究所

政所里佳.(2018).「パラフレージング・レベルによる錯乱肢の作成ー語答分析に基づくテキスト理解度の診断に向けてー」. *EIKEN BULLETIN, 30,* 85-108.

水本篤. (2024). AI のある英語教育・研究 Let There Be AI!. *KELES* ジャーナル, 9, 52-58.

Mizumoto, A. (2025). Langtest. https://langtest.jp/ (2025年7月22日閲覧)

文部科学省 初等中等教育局.(2024.12).「生成 AI の利活用に関するガイドライン ver.2.0」. 文部科学省.

文部科学省.(2018).「中学校学習指導要領(平成29年告示)」. 文部科学省.

日本英語検定協会.(2024).『英検2級一次試験過去問』. https://www.eiken.or.jp/eiken/exam/grade_2/(2025年7月22日閲覧)

日本リサーチセンター.(2025.7.7).「【NRC デイリートラッキング】生成 AI の利用経験　2025年6月調査」. https://www.nrc.co.jp/report/250707.html (2025年8月28日閲覧)

根岸雅史.(2017).『テストが導く英語教育改革 「無責任なテスト」への処方箋』. 三省堂.

OpenAI. (2024). ChatGPT- 4o(2024.5.31) [Large language model]. https://ChatGPT.com/

Oxford University Press. (2020.3.10). *Oxford Advanced Learner's Dictionary 10th edition*

R Core Team (2025). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Sato, T. & TeachLex Scope Team. (2025). TeachLex. https://teachlex.pythonanywhere.com/

清水真紀.(2005).「リーディングテストにおける質問タイプーパラフレーズ・推論・テーマ質問と処理レベルの観点からー」. *EIKEN BULLETIN, 17,* 48-62.

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology, 27*(3), 27-40.

竹内理・水本篤.(2023).「外国語教育研究ハンドブック―研究手法のより良い理解のために―」. 松柏社.

Terao, T., & Ishii, H. (2020). A comparison of distractor selection among proficiency levels in reading tests: A focus on summarization processes in Japanese EFL learners. *Sage Open, 10* (1), 2158244020902087.

卯城祐司.(2011).『英語で英語を読む授業』. 研究社.

Ushiro, Y, Morimoto, Y., Hijikata, Y., Nakagawa, C., Watanabe, F., Kai, A., Shimizu, H., Koga, T., Ohno, M., Umehara, C., Hamano, Y., Narumi, T., Tao, N., Shimada, S., Kiyama, M., Suzuki, R., Kurisu, Y., & Gomi, N. (2007). What Makes Distractors Plausible in Multiple-choice Reading Tests? *Japanese Language Testing Association, 10,* 56-67. https://doi.org/10.20622/jltaj.10.0_56

Ushiro, Y., Nakagawa, C., Morimoto, Y., Hijikata, Y., Watanabe, F., & Kai, A. (2008). Effects of question types on item difficulty in two reading test formats: Open-ended and multiple-choice. *ARELE: Annual Review of English Language Education in Japan, 19,* 201-210.

若林俊輔・根岸雅史.(1993).『無責任なテストが「落ちこぼれ」を作る』. 大修館書店.

Watanabe, T., & Saida, C. (2024). ChatGPT's Performance on the English Reading Sections of the Common Test for University Admissions in Japan. *KATE Journal, 38,* 155-168.

資料1：オリジナルとパラフレーズの正答選択肢一覧

項目番号			選択肢	総語数
A	(1)	オリジナル	will take part in an event at a university in Petersburg.	11
		パラフレーズ	will join an event at a university in Petersburg.	9
	(2)	オリジナル	Hand in some forms at least seven days before then.	10
		パラフレーズ	Make sure to give the forms in at least seven days ahead of time.	14
	(3)	オリジナル	Work with two other employees to prepare for February 24.	10
		パラフレーズ	Team up with two other employees to get ready for February 24.	12
B	(4)	オリジナル	Keep an eye on how words are used in the English language.	12
		パラフレーズ	Notice how words are used when speaking or writing in English.	11
	(5)	オリジナル	a large number of people are using the word in the same way.	13
		パラフレーズ	a lot of people are using the word in a similar way.	12
	(6)	オリジナル	has gained a different meaning from its original one.	9
		パラフレーズ	has changed from what it was originally.	7
	(7)	オリジナル	To make the final decision about changes to the content of the dictionary.	13
		パラフレーズ	To make the last decision about updating the dictionary content.	10
C	(8)	オリジナル	She received Nobel Prizes in more than one subject area.	10
		パラフレーズ	She won Nobel Prizes in different subject areas.	8
	(9)	オリジナル	the university did not allow women to become students.	9
		パラフレーズ	women were not allowed to be students at the university.	10
	(10)	オリジナル	The food she ate was not good enough for her to stay healthy.	13
		パラフレーズ	She did not eat food that was healthy enough for her.	11
	(11)	オリジナル	worked at the same university where Marie Curie was doing research.	11
		パラフレーズ	worked at the same university where Marie Curie was conducting her research.	12
	(12)	オリジナル	Marie Curie and her sister agreed to help each other to study at university.	14
		パラフレーズ	Marie Curie and her sister promised to assist each other in their university education.	14
D	(13)	オリジナル	made some recommendations to him and his friend.	8
		パラフレーズ	gave them a few recommendations to consider.	7
	(14)	オリジナル	The way that it looked was very appealing	8
		パラフレーズ	Its appearance was very attractive.	5
	(15)	オリジナル	to talk with her about working at her restaurant.	9
		パラフレーズ	to have a conversation with her about a job at her restaurant.	12
E	(16)	オリジナル	They left their farms and started buying things that were made in factories.	13
		パラフレーズ	They moved away from their farms and began purchasing factory-made goods.	11
	(17)	オリジナル	The lives of people would improve if they returned to making things by hand.	14
		パラフレーズ	If people returned to handmade production, their lives would become better.	11
	(18)	オリジナル	did not have enough involvement in the production process.	9
		パラフレーズ	lacked sufficient involvement in how things were produced.	8
	(19)	オリジナル	the products were expensive and not easily available.	8
		パラフレーズ	the goods were expensive and difficult to access.	8
	(20)	オリジナル	Technology developed in the nineteenth century had a significant effect on UK society.	13
		パラフレーズ	The UK society was strongly affected by the technology created in the nineteenth century.	14
F	(21)	オリジナル	To congratulate her on her first year as a member of the fitness center.	14
		パラフレーズ	To give her best wishes for her first anniversary at the fitness center.	13
	(22)	オリジナル	arrange to be in a feature in an online magazine.	10
		パラフレーズ	plan to be included in an article in an online magazine.	11
	(23)	オリジナル	people who are introduced by other members.	7
		パラフレーズ	people who are brought in by existing members.	8

(資料1 続き)

項目番号		選択肢		総語数
G	(24)	オリジナル	The location where they live is only in a certain area of Africa.	13
		パラフレーズ	Their home is found only in a certain region of Africa.	11
	(25)	オリジナル	bonobos were seen to help other bonobos without getting a reward.	10
		パラフレーズ	bonobos helped each other without expecting anything in return.	9
	(26)	オリジナル	bonobos could feel others' feelings even more than humans.	9
		パラフレーズ	bonobos were able to understand others' emotions even better than humans.	11
	(27)	オリジナル	tend to be less friendly because there is a strong leader among them.	13
		パラフレーズ	tend to show less friendliness because a dominant leader is among them.	12
	(28)	オリジナル	Bonobos live peacefully in groups that have no strong leaders.	10
		パラフレーズ	Bonobos stay in harmony in groups where no strong leader is present.	12

(注)文頭が小文字で始まっている選択肢は問題文の続きを選ぶ形の問題のため,原文のまま載せている。