第33回 研究助成

A 研究部門 ● 報告Ⅰ ● 英語能力テストに関する研究

読み手の理解度を考慮した英作文の正確性指標の量的妥当化:どのような誤りが"重い"のか

研究代表者:茨城県／筑波大学大学院　在籍　岡　秀亮
共同研究者:茨城県／筑波大学大学院　在籍　加藤 剛史

《研究助言者:寺内　一》

概要

本研究の目的は,(1)英作文における書き手の誤りと,それらが読み手の理解を阻害する度合い(重み付けされた正確性指標:WCR)を調査し,(2)熟達度に応じて,誤りとWCRがどのように変化するのかを明らかにすることである。WCRの長所は,誤りの種類を読み手の理解を阻害する度合いに応じて3段階に同定することで,教師によるフィードバックの与え方や指導方法をより明確にすることを可能にする点にある。一方,短所として,当該指標の評定には評定者の「理解の阻害度」が影響するため,高い信頼性を得ることが難しい。そこで,本研究は,Polio and Shea(2014)にみられる25の文法的誤りがその種類に応じてどのように重みづけされるべきなのか,対応分析を用いて検討した。その結果,3つのカテゴリーが得られ,各カテゴリーで特徴的な誤りの種類を特定することができた。さらに,熟達度間におけるWCRの変化を分析した結果,軽微な誤りは熟達度が上昇するにつれて減少する一方で,読み手への影響度が大きい誤りの数は,熟達度が上がると増加する傾向にあることがわかった。

1 はじめに

英語を正確に書く能力(正確性)は,書類やメールを英語で書く必要がある国際社会や,研究で得られた知見を論文という形で広める学術分野において,必要不可欠な能力の1つである。英語を正確に書く能力を身につけることで,世界中の人々に対して効果的に自分の考えを伝えることができる。

教育現場において,英語を正確に書くことができる学習者を育成するには,正確性を適切に評価できる測定方法の確立が必要不可欠である。なぜなら,学習者が持っている能力の程度や発達段階を正確に把握した上で,適切な指導(e.g.,訂正フィードバック)を選択・実施することで,効率的に正確性を向上させることができるからである。

近年,英作文の正確性をより詳細に測定・評価することができる正確性指標(i.e., weighted clause ratio; WCR)が提唱された(Foster & Wigglesworth, 2016)。WCRを用いることで,読み手の理解度に与える影響度に応じた評価を行うことが可能となるため,正確性をより詳細に測定することができる。さらに,縦断的・横断的な調査を行うことで,正確性のより細かな発達を可視化することができる。

しかしながら,WCRによる正確性の測定には,評価尺度の曖昧さが指摘されている(e.g., Oka, 2021)。そこで本研究は,近年提唱された新しい正確性指標であるWCRを用いて,(1)WCRによる正確性測定の評価方法の確立を目指す。さらに,(2)確立された評価尺度を用いて,正確性の発達過程を明らかにする。本研究を行うことで,WCRを用いた信頼性の高い測定方法を提唱する

ことができるため，教員による学習者の正確性測定の質を担保することが可能となる。さらに，WCRを用いた際の正確性の発達を明らかにすることで，学習者の発達段階に応じた英作文指導を行うことが可能になる。

2 先行研究

2.1 英作文における正確性の測定

正確性（accuracy）は，“the ability to produce the target-like and error-free languages”と定義されている（Housen, Kuiken, & Vedder, 2012, p.2）。また，正確性は，ライティング熟達度を構成する3つの構成概念（complexity: 複雑性，accuracy: 正確性，fluency: 流暢性）の1つとされており（Housen & Kuiken, 2009），多くの研究で使用されてきた（e.g., Kuiken & Vedder, 2008）。CAFには，各構成概念を反映すると想定される指標（index）があり，研究者はその指標値を使用して，学習者が持っている能力の程度を推測したり，指導法の効果を検証する。

英作文における正確性の測定には，これまで非常に多くの指標が提案されてきた。Wolfe-Quintero, Inagaki, and Kim（1998）は，第二言語（second language: L2）ライティングを対象とした研究で使用された正確性指標の妥当性を大規模に調査した。その結果，誤りのないT-unit（i.e., 独立節とそれにつながる従属節からなる主節を含む単位）の総数やerror-free T-units per all T-units（EFTR）が，正確性を測定するために最適な指標であると結論づけた。しかしながら，誤りの数を数えた指標（numbers of errors; Chandler, 2003），エラーのない節の数（number of error-free clauses）を計算する（e.g., Ellis & Yuan, 2004）など，研究間で異なった指標が用いられ続けた。

Wolfe-Quintero et al.（1998）の研究から約20年間，様々な正確性指標が用いられてきた一方で，近年，CAFの測定方法を疑問視する動きが見られている。例えば，Housen et al.（2012）は，CAFをどのように測定すべきかについて一致した見解が得られていないことを指摘し，正確な測定結果を得るためには妥当性・信頼性を確立する必要があると主張している。これをきっかけとし，複雑性の因子的妥当性（Kato, 2019）や言語発達との関連性（e.g., Kyle & Crossley, 2018）に関する研究が多く行われ始めた。

正確性に関しては，正確性を適切に測定するための指標はまだ不明確である（Michel, 2017）という指摘があるが，正確性測定に関する研究が進んできている。Foster & Wigglesworth（2016）は，誤りのない節を単位とする指標（e.g., error-free clause per clauses: EFCR）は，正確性の測定に有効な指標の1つであると主張している。その理由として，（1）節の定義がしやすいことや，（2）T-unitの長さと比較して節は短いことから，英作文データを細かく分析することができるためとされている。

しかしながら，Foster & Wigglesworthは，先行研究での批判と同様に（e.g., Kuiken & Vedder, 2008; Polio, 1997），誤りのない節を単位とした指標を正確性の測定に用いることについて，依然として課題点があると指摘している。EFCRのような指標を用いた場合，非常に軽微な誤り（主語と動詞の一致の誤りなど）を持つ節と，重大な誤り（重大な単語選択の誤りなど）を持つ節は同じ得点が与えられる。ゆえに，Evans, Hartshorn, Cox, & de Jel.（2014）も指摘するように，EFCRなどを使用すると正確性の潜在的な違いが見落とされる可能性があり，指標の信頼性と妥当性が損なわれる可能性があると主張している。

2.2 英作文の正確性を測る指標としてのweighted clause ratio

Foster & Wigglesworth（2016）は，誤りのない節を単位とする正確性指標の限界点を踏まえ，新しい正確性指標としてWCRを提案した。WCRは，“the degree to which a learners’ performance is more or less successful in achieving the task’s goals efficiently”（Pallotti, 2009, p.7）というadequacyの概念に基づいて，正確性を測定するために開発された。

WCRの評価尺度は，「4つのカテゴリー」，「カテゴリーの定義」，「得点」の3つで構成されてい

る（表1）。WCRを用いて評価を行う際は，英作文の全文を節に分割した後，定義に基づいて各節をいずれかのカテゴリーに分類し，各カテゴリーに応じた得点が節に与えられる。なお，レベル3の文節であってもある程度は言語的に正確であるため，0をつけるべきではないとされている（Foster & Wigglesworth, 2016）。

■表1: WCRの評価尺度（筆者訳）

カテゴリー	定義	得点
Entirely accurate	誤りがない節	1.0
Level 1	意図された意味を曲げない程度の誤り（e.g., 三単現のs）を含む節	0.8
Level 2	意図された意味の理解が（常にではないが）困難な誤り（e.g., 語彙の選択や語順）を含んだ節	0.5
Level 3	意図された意味の理解が不可能な誤りを含んだ節	0.1

EFCRのような「正確か不正確か」という2値的な評定とは異なり，WCRは節内に含まれる誤りをその影響度に応じて段階評価するという特徴を持っている。WCRの評価尺度は4つのカテゴリーから構成されているため，ライティングパフォーマンスにおける正確性のわずかな変化に敏感である。そのため，研究者は，細かな違いを見過ごすことなく，正確性を測定・評価することができる。さらに，ライティングパフォーマンスにおける正確性の小さな変化を経時的に追跡することができる（Evans et al., 2014）。一方で，教師は，あらゆる種類の誤りに対処する指導を実施する代わりに，読者の理解に大きく影響する誤りに焦点を当てることができる。ゆえに，学習者は，読み手を意識しながら自分が意図したメッセージを書き直すことができる。

2.3 WCRによる正確性測定の信頼性・妥当性について

WCRは近年提唱された指標であるため，指標の妥当性や信頼性に焦点が置かれた研究が多い（e.g., Evans et al., 2014; Oka, 2021; Polio & Shea, 2014）。Evans et al.（2014）は，多相Raschモデルを用いて，ESL学習者のライティングパフォーマンスを対象に，3つの正確性指標（EFCR, EFTR, WCR）の評価の信頼性を検証した。調査の結果，指標値はライティングのトピックの違いに影響されるものの，これらの指標の信頼性は高いと結論づけられた。

また，WCRとは少し異なる指標（Weighted error-free T-units）が用いられているが，Polio and Shea（2014）も，ESL学習者が1学期の異なる時期に書いた英作文を用いて，正確性測定の信頼性と妥当性を探った。彼らは，Weighted error-free T-units と他の評価方法の関連性を調査するために，全体的尺度（正確性を対象としたものに限定），EFCR，特定の誤りに対する正確性指標を用いた。その結果，WCRと同様Weighted error-free T-unitsの評価者間信頼性は高い（r= 0.84）ことがわかった。

ESL環境においては，WCRを用いた評価の信頼性がある程度担保されていることが示された一方で，EFL学習者を取り上げた研究は行われていなかった。そこで，Oka（2021）は，日本人EFL学習者が作成した英作文を対象に，WCRによる評価の信頼性を検証した。Okaの研究では，グローバルエラーやローカルエラーに関わらず，日本人EFL学習者は様々な誤り（e.g., 動詞の選択や相対節）を産出する傾向があるというコーパス研究（e.g., Abe, 2019; Nagata et al., 2011）の知見から，EFL学習者に対する現行のWCRの尺度を見直す必要性を主張した。そこで，Polio and Shea（2014）の誤り表を参考に，カテゴリーに属する誤りの種類を詳細にした評価尺度を作成し（資料1），一般化可能性理論を用いて評価の信頼性を調査した。その結果，評価尺度の信頼性は非常に高いことが示された（G coefficient = .91）。

3 研究課題

WCRの評価尺度に関する研究は徐々に進んできているものの, 未だ実証的な分析に基づいた評価尺度は得られていない。Oka(2021)では, 評価全体の信頼性は高いことが示された一方で, 各カテゴリーと誤りの種類の一致は評価者間の協議のみで決まっており, 実証的な分析を経て得られたわけではない。

そこで, 本研究は, WCRの併存的妥当性を検証する目的で, WCRにおける各カテゴリーと関係性のある誤りの種類が, 先行研究で指摘されているようになるのかを検証する。その後, 熟達度が上昇するにつれて, WCRの各カテゴリーがどのように変化するのかを解明する。これらを明らかにするために, 2つの研究課題(Research questions: RQs)を設定した。

RQ1 日本人英語学習者が作成した英作文を対象に, WCRを用いて評価したとき, どの誤りがWCRのLv.1, Lv.2, Lv.3の各段階において特徴的であるのか

RQ2 日本人英語学習者が作成した英作文を対象に, WCRを用いて評価したとき, 各段階に属する節は熟達度間でどのように変化するのか

4 研究方法

4.1 データ

本研究は, International Corpus Network of Asian Learners of English(ICNALE)コーパス(Ishikawa, 2013)に収録されている英作文データを使用する。このコーパスに含まれている英作文は, アジアの国と地域(e.g., 中国や日本, フィリピン)にある2800名の大学生によって産出された。また, ICNALEでは, 書き手の熟達度にCEFRレベルが付与されている。ICNALEプロジェクトに参加した学習者は, TOEFL, TOEIC, IELTSなどによるスコアを提出したのち, 語彙テストを受験した(Nation & Beglar, 2007)。参加者によって報告された熟達度と語彙テストのスコアを使用し, 4つのCEFRレベル(A2, B1_1, B1_2, B2+)に分類された。A2(N = 154)は一番低い熟達度, B1_1(N = 179)はB1の中でも低い熟達度, B1_2(N= 49)はB1の中でも高い熟達度, B2+(N = 18)はB2およびC1とC2を含めた熟達度になっている(Ishikawa, 2013)。

ICNALEで得られたデータは, 以下のような細かな統制のもとで得られた。英作文は, 2つの異なるトピックが使用された。1つは, "It is important for college students to have a part-time job(PTJ)" であり, もう1つは "Smoking should be completely banned at all the restaurants in the country(SMK)" である。参加者は1つの英作文につき30分で書くように指示された。また, 英作文作成時, 参加者は辞書を使うことができなかったが, スペルチェッカーのみは使用が許されていた。また, 語数は200字から300字に制限されていることで, 語数による影響を最小限に抑えられるようになっている。

本研究では, 日本人英語学習者によって書かれた英作文(SMK)400個のうち, 100個を抽出し分析対象とする。A2から25人, B1_1から25人, B1_2から32人, B2+から18人をランダムに抽出した。参加者は経営学や医学など幅広い分野にわたっていた。

4.2 WCRの評価尺度

本研究は, Oka(2021)で使用された評価尺度を用いた。英語教育を専攻する日本語母語話者3名に対し評定者トレーニングを行った上で, 著者を含めた4名の評価者が, 英作文中の各節にWCRの評点を付与した。その際, 節中に文法上の誤りが認められた場合には, Polio and Shea (2014)が提案する25種類のエラー分類に基づき, エラータグを付与する。複数回の修正が行われた上で, 最終的なWCRの評価尺度が完成した (資料1)。この評価尺度による評価は, 信頼性が高いと示された。

4.3 採点

採点の前に，調査者が各英作文の文章を節に分割し，3名の評価者が全節をチェックした。その後，最終版の評価尺度を用いて，4評価者がすべての英作文を独立して評価した。評価者は，評価者トレーニングと同様の手順で，各文節の誤りを見つけ，その誤りが読者の理解にどの程度影響を与えるかによって，誤りの重大度を評価することが求められた。なお，同じ誤り（単語の誤りなど）であっても，その重大性は文脈に左右されることが多いため，異なるレベルに分類されることがあった。文脈に応じて分類されるからである。評価者は，このような誤りを見つけた場合，評価尺度の定義を読んで分類することに同意した。さらに，同じ節に複数の誤り（Lv.1とLv.3の誤りなど）がある場合は，Foster and Wigglesworth（2016）が提案しているように，影響度が大きいレベルの誤りに応じて節を分類した。最終的なWCRスコアは以下のように計算した［WCR＝（正確な節の数×1.0＋Lv.1の節の数×0.8＋Lv.2の節の数×0.5＋Lv.3の節の数×0.1）/ エッセイの全節］。

4.4 エラータグの付け方

本研究で使用するデータは，Polio and Shea（2014）の誤り表を用いてタグづけされた。また，日本人英語学習者が多く産出する誤り（e.g.，接続詞）などが追加された。誤りタグは表2で示されたように行われた。

■表2: 誤りタグ

エラーのカテゴリー	タグ	例
冠詞	art	<art crr="a"> </art> restaurant is…
主語と動詞の一致	sv_agree	There are many things which <sv_agree crr="are">is</sv_agree>…
名詞の単数 / 複数	sg_plu	I hate <sg_plu crr="smokers">smoker</sg_plu>.

あるべき言語項目が書かれていない場合は（e.g.,冠詞の例），欠如している箇所に誤りタグを付与した。また，書かれている言語項目や表現が間違っている場合は，誤っている箇所の前には誤りの種類（e.g., sv_agree）と正しい文法や形（i.e., crr=""）を追加し，後には指摘した誤りの終わりを示す記号（/）と誤りの種類を追加した。

4.5 分析手法

まず，WCRの各段階と各誤りの間の関連の強さを算出するために，対応分析を行った。対応分析を行う際は，投野・望月（2012）を参考に，「WCRの段階」×「26タイプの誤り」のクロス集計表を作成した。本研究では，近年統計解析に使用され始めているフリーソフトウェアであるRを使用する。統計解析を行うためのパッケージをダウンロードし、その中にある関数を使用して分析を実行する。対応分析を行うために，CA関数（Lê, Josse, & Husson, 2008）を使用して，分析を実行した。

さらに，カイ二乗検定を使用し，WCRにおける各カテゴリーに含まれるエラー頻度を熟達度別に検証した。カイ二乗検定は，人数や頻度の差の比較に用いられる統計手法の1つである。本研究では、Rにおけるchi関数を使用して分析を実行した。

5 結果と考察

5.1 WCRの評価尺度の量的妥当性（RQ1）

RQ1では，WCRの評価尺度における各段階（i.e., Lv.1, Lv.2, Lv.3）で特徴的な誤りを特定するために，対応分析を行った。このRQに取り組むことで，評価者は誤りを含んだ節をより正確に分類できるため，WCRの評価をより信頼性の高いものにすることができる。また，Oka（2021）で示された評価尺度（資料1）において，カテゴリーの判別とはあまり関係のない誤りも特定することが

できるため，評価のpracticality（e.g., Bachman & Palmer, 1996）を改善することもできる。

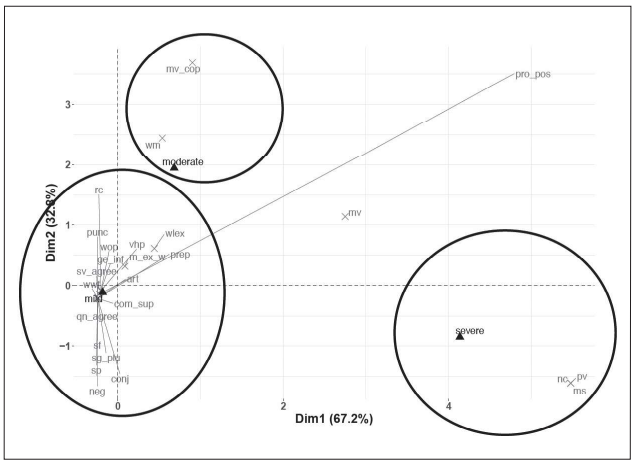
表3は，1万語あたりにおける誤りの頻度をクロス集計でまとめたものである。日本人英語学習者（N = 100）が作成した英作文を対象に誤りの数を集計し，1万語あたりの誤りの頻度を算出した結果，様々な誤りが出現していることがわかった。特に，Lv.1に該当する誤り（e.g., 冠詞）の数が非常に多いことがわかった。一方で，Lv.2に該当する誤り（e.g., 語彙）やLv.3に該当する誤り（e.g., 意味が読み取れない）の数は少なく，その種類も限定的であることが示された。

表3: 1万語あたりの誤りの頻度

誤りの種類	タグ	Lv.1	Lv.2	Lv.3
冠詞	art	244.20	0.45	0.00
比較級・最上級	com_sup	1.35	0.00	0.00
接続詞	conj	148.05	0.00	0.00
不定詞・動名詞	ge_inf	24.37	0.90	0.00
語彙	m_ex_w	132.26	23.47	4.51
主語の欠如	ms	0.00	0.00	4.06
動詞の欠如	mv	0.90	5.87	4.97
動詞の目的語の欠如・余剰	mv_cop	0.00	7.67	0.00
意味が読み取れない	nc	0.00	0.00	8.12
否定	neg	1.35	0.00	0.00
前置詞	prep	130.45	0.00	0.00
所有・代名詞	pro_pos	24.83	0.45	0.00
修辞法	punc	181.91	0.00	0.00
受動態	pv	0.00	0.00	10.38
数詞	qn_agree	0.90	0.00	0.00
関係詞・関係副詞	rc	12.64	0.45	0.00
断片文	sf	2.26	0.00	0.00
名詞の単数 / 複数	sg_plu	185.52	0.00	0.00
スペリング	sp	19.86	0.00	0.00
主語ー動詞の一致	sv_agree	43.33	0.00	0.00
動詞の時制・分詞	vhp	27.08	0.90	0.00
語彙の選択	wlex	133.61	46.04	13.99
法助動詞	wm	6.32	13.54	0.00
語順	wop	24.83	0.90	0.00
品詞の形	wwf	54.17	0.00	0.00

注. Lv.1 = Level 1; Lv.2 = Level 2; Lv.3 = Level 3.

さらに，WCRの評価尺度における各カテゴリーに当てはまる誤りの種類を特定するために，上記の結果を使用して対応分析を行った。その結果が，図1と図2に示されている。図1にある黒い三角で示された部分に近いほど，誤りはそのカテゴリーに含まれると解釈することができる。また，WCRの評価尺度におけるカテゴリーがどの程度の影響度かを示すために，Lv.1のカテゴリーをmild, Lv.2のカテゴリーをmoderate, Lv.3のカテゴリーをsevereとした。また，図2は，対応分析の結果をヒートマップで示したものである。誤りの種類は右に，影響度の大きさは下に示されている。色が黒になるほど，カテゴリーとの関連性が強いと解釈することができる。



■図1: 散布図

■図2: ヒートマップ

分析の結果，多くの誤りが，Lv.1のカテゴリーの付近（左下の円）に集まっていることが示された。また，ヒートマップにおいてもMildのカテゴリーにおける黒の割合が多いことからも，同様のことが言える。つまり，誤り表にあるほとんどのエラーが，Lv.1のカテゴリーに含まれる傾向にあると言える。これは，本研究が使用した英作文には，三単現のsや動名詞の誤りが多く出現していたことが要因であると言える。また，Oka（2021）で示された誤り表において，Lv.1に該当する誤りが一番多いことも要因の1つと考えられる。また，Lv.1に関する判断において，関連性の低い誤りは見られなかった。

また，Lv.2のカテゴリー（左上の円）には，「動詞の目的語の欠如・余剰（mv_cop）」や「法助動詞（wm）」が強く関連していることが示された。また，ヒートマップにおいても，2つのエラーがModerateの部分にて黒く表示されている。ゆえに，これらの誤りは，評価者が節をLv.2のカテゴリーに分類をする上で，中心となる誤りであると言える。Lv.2のカテゴリーは，読み手の理解度を大きく阻害するカテゴリーである（Foster & Wigglesworth, 2016）。法助動詞や動詞の目的語は，英作文にある節の内容を理解する上で，重要な言語項目である。例えば，“If I knew<mv_cop crr=“the fact”> </mv_cop> then”のように，動詞の目的語が欠如している場合，何を知っているのかが読み手には伝わりにくい。また，法助動詞の誤りは，学習者が持っている意見に対する確信度や仮定法過去などにおいて誤りが多く，文脈によっては読み手の理解度を大きく阻害する言語項目であると言える。

一方で，誤り表にあるその他の誤りは，Lv.2のカテゴリーとは関連性が低いことも示された。Lv.2のカテゴリーには，冠詞や動名詞の誤りも含まれているが，これらの誤りは，読み手の理解度にほとんど影響を与えないLv.1のカテゴリーであると判断されることが多いためだと言える。これは，誤りの頻度数を集計した表3からも言える。また，語彙の誤り（e.g., m/ex_w, wlex）は全ての段階に入っているが，本研究では，特にLv.1と判断される誤りが多かったため，Lv.2のカテゴリーには含まれなかったと言える。

さらに，読み手の理解度に大きく影響を及ぼすLv.3のカテゴリー（右下の円）には，「意味が読み取れない（nc）」誤りや「主語の欠如（ms）」，「受動態（pv）」に関する誤りが含まれることが示された。ヒートマップにおいても，同様のことが示されている。特に，主語の欠如には，itのような代名詞の場合や人称代名詞の場合が考えられるが，どのような場面においても，主語がなければ節の内容を理解することは困難であると言える。

対応分析の結果，誤り表のうち，ほぼ全ての誤りがいずれかのカテゴリーに含まれることがわかった。しかし，「一般動詞の欠如に関する誤り（mv）」のみ，Lv.2とLv.3のカテゴリーの中間部分に位置していることが明らかとなった。原因として，本研究が使用した英作文には，文脈によってはLv.2に分類される場合もあれば，Lv.3であると判断される場合もあったためと考えられる。

以上の結果をまとめると，各カテゴリーに含まれる誤りがエラー表に示されているようにまとまっていることがわかった。ゆえに，Oka(2021)において評価者間の協議によって作成された評価尺度は，エラーを含んだ節を分類する上で，信頼性のある評価尺度であると言える。

5.2 WCRにおける各段階の発達 (RQ2)

RQ2は，熟達度が上がるにつれて，各カテゴリーに含まれる誤りの頻度がどのように変化するのかを明らかにすることが目的であった。誤りの種類とWCRの熟達度に応じた変化を可視化することで，誤りの単純な頻度だけでなくWCRの各段階がどのような過程で減少するのかを熟達度別に明らかにできる。

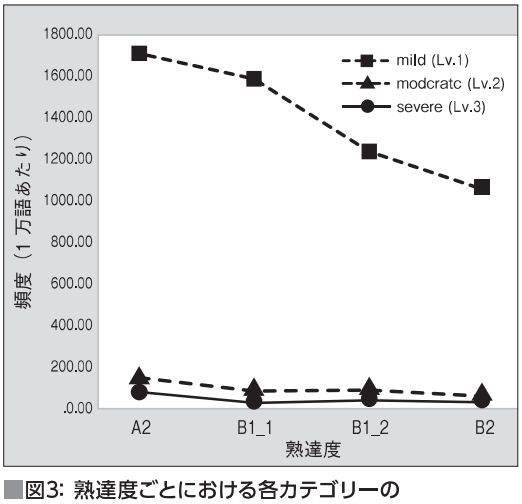
表4と図3は，各カテゴリーに属する誤りの頻度（1万語あたり）を熟達度別に示したものである。各カテゴリーに属する誤りの全体的な頻度が，熟達度ごとに変化しているかどうかを調べるために，χ^2検定を行った。分析の結果，Lv.1のカテゴリーにおいて，A2とB1_1間で有意な差が見られた，[$\chi^2(1) = 4.53, p = 0.03$)]。また，A2とB1_2間 [$\chi^2(1) = 76.8, p < .001$)]，A2とB2+間 [$\chi^2(1) = 151.6, p < .001$)]，B1_1とB1_2間 [$\chi^2(1) = 44.2, p < .001$)]，B1_1とB2+間 [$\chi^2(1) = 104.0, p < .001$)]，B1_2とB2+間 [$\chi^2(1) = 12.8, p < .001$)] で有意な差が見られた。

■表4: 熟達度ごとにおける各カテゴリーのエラー頻度数

カテゴリー	A2	B1_1	B1_2	B2+
mild(Lv.1)	1710.40	1588.13	1234.82	1063.12
moderate(Lv.2)	143.65	89.05	102.33	59.33
severe(Lv.3)	86.19	27.83	40.93	28.48

以上の結果から，Lv.1のカテゴリーにおいては，誤りの数が全ての熟達度間で有意に減少することが示された。つまり，熟達度が上昇するに従って，読み手の影響度にほとんど影響を与えない軽微な誤りは，徐々に減少していく傾向にあると言える。Lv.1に含まれる誤りは，三単現のsや冠詞，接続詞に関する誤りなど，英語初学者に多く見られる誤りであるため，熟達度が上昇するにつれて，このような誤りは減少すると言える。ライティングパフォーマンスとスピーキングパフォーマンスに含まれる誤りの数を，コーパスを使用して調査したAbe(2019)においても，三単現のsは熟達度が上がるにつれて減少していることが示されている。

次に，Lv.2のカテゴリーにおけるエラーの全体的な頻度さを熟達度ごとに比較した。その結果，A2とB1_1間で有意な差が見られた，$\chi^2(1) = 12.8, p < .001$)。また，A2とB1_2間 [$\chi^2(1) = 6.94, p = .008$)]，A2とB2+間 [$\chi^2(1) = 35.0, p < .001$)]，B1_1とB2+間 [$\chi^2(1) = 6.0, p = .001$)]，B1_2とB2+間 [$\chi^2(1) = 11.4, p < .001$)] で有意な差が見られた。しかし，B1_1とB1_2間では有意な差が見られなかった，$\chi^2(1) = 0.92, p = .33$。



■図3: 熟達度ごとにおける各カテゴリーのエラー頻度数の推移

以上の分析から，Lv.2のカテゴリーにおいては，ほぼ全ての熟達度段階で有意に減少することがわかった一方で，B1_1とB1_2間では有意な

差が見られなかった。逆に，B1_1とB1_2間では統計的な有意差はないものの，Lv.2に該当する誤りの数が上昇していることが示された。これは，B1_2の学習者は，B1_1の学習者よりも複雑な文構造を使用したり，より低頻度な英語を使用したことにより，誤りの数が上昇した可能性がある。

最後に，Lv.3のカテゴリーにおけるエラーの全体的な頻度を熟達度ごとに比較した。その結果，Lv.1のカテゴリーにおいて，A2とB1_1間で有意な差が見られた，$\chi^2(1) = 29.9$, $p < .001$。また，A2とB1_2間 [$\chi^2(1) = 16.11$, $p < .001$]，A2とB2+間 [$\chi^2(1) = 29.0$, $p < .001$] においても有意な差が見られた。一方で，B1_1とB1_2間 [$\chi^2(1) = 2.50$, $p = .11$)]，B1_1とB2+間 [$\chi^2(1) = 0.0$, $p = .93$)]，B1_2とB2+間 [$\chi^2(1) = 1.87$, $p = .17$)] では有意な差は見られなかった。

分析の結果から，比較的，熟達度が低い学習者（A2）がLv.3のカテゴリーを多く産出することがわかった。読み手による理解が不可能な誤りは，初学者が多く産出する誤りが含まれているため，A2で多く見られたと言える。このことが，熟達度が高い学習者（B1_1, B1_2, B2+）では，Lv.3に該当する誤りは産出されているものの，有意な差が見られていない要因であると考えられる。しかし，Lv.2のカテゴリーの場合と同様に，B1_1とB1_2間では統計的な有意差はないものの，Lv.3に該当する誤りの数が上昇していることが示された。

6 結論と今後の課題

Summary of findings

本研究は，読み手の理解度に与える影響度に応じた正確性評価を行うことができるWCRに注目し，その評価尺度の併存的妥当性を検証した（RQ1）。具体的には，対応分析の結果で得られたグルーピングが，Oka（2021）で示されている評価尺度と同様の結果になるかどうかを調査した。対応分析の結果，Oka（2021）で示された評価尺度と同様に，3つのグループを得ることができた。さらに，本研究は，WCRにおける各カテゴリー（Lv.1, Lv.2, Lv.3）が学習者の熟達度に応じてどのように減少するのかを明らかにした（RQ2）。分析の結果，Lv.1のカテゴリーは全熟達度間で有意に減少したことがわかった一方で，Lv.2とLv.3のカテゴリーは熟達度が上昇したとしても，減少しない区間（B1_1とB2+）や逆に誤りの数が増える熟達度区間（B1_1とB1_2）があることが示された。

Implications for researchers and educators

本研究は，コミュニケーションを意識した正確性評価に着目し，WCRを英作文の正確性評価に応用した。WCRを用いた正確性評価を行うことで，教師は読み手を意識したライティング指導と評価の実施が可能になる。さらに，信頼性の高い評価尺度を提供することで，研究者は正確性の適切な測定，及び正確性の詳細な発達過程を解明することができる。

本研究によって得られた教育的示唆は2点ある。1点目は，高い信頼性を担保したWCRの評価尺度を示すことができた点である。対応分析の結果，WCRにおける各カテゴリーと関連性の高い誤りの種類を特定することができ，Oka（2021）が示した評価尺度と非常に類似していることがわかった。これらのことから，教師は学習者が持っている正確性を，従来から使用されている指標よりも詳細に，かつ正確に把握することができる。

2点目は，ライティング指導の時間を費やすべきカテゴリーとその熟達度帯を示すことができた点である。分析の結果，Lv.1に該当する誤りは，熟達度が上がるにつれてその数は減少していくことが示された。一方で，Lv.2のカテゴリーの誤りの数とLv.3のカテゴリーの誤りは熟達度が上昇してもその数は減少せず，むしろ増加する区間もあることがわかった。これらのことから，Lv.1のような軽微な誤りに対しては，ライティング能力全体を向上させる指導を継続的に行うことが有効だと考えられる。一方で，Lv.2のカテゴリーとLv.3のカテゴリーにおいては，熟達度が中位程度の学習者（B1_1とB1_2）において，誤りに対する指導の時間を増やす必要があると考えられる。

Limitations

本研究は，英作文における正確性の評価，及び正確性の発達を明らかにすることができたものの，3点の課題点がある。1点目は，評価者が英語母語話者ではないことである。本研究はOka (2021) で使用されたデータを使用したが，評価者はEFL及びESL学習者であった。英語母語話者がWCRを用いて正確性を評価した場合，誤りの影響度や分類が異なる可能性がある。今後は，英語母語話者による評価とESL・EFL学習者による評価を比較・検討する必要がある。

2点目は，誤りの種類が限定的である点である。本研究では，評価にかかる時間やpracticalityを考慮し，Polio and Shea (2014) が作成した誤り表をもとにタグづけされたデータを用いた。しかし，より細かく誤りタグをつけて正確性の発達を検証している研究もある (e.g., Thewissen, 2013)。今後は，タグづけにかかる時間や分類の適切さを考慮した誤り表を検討する必要がある。

3点目は，異なる学校種に所属する学習者が作成した英作文を分析対象にできなかった点である。例えば，日本人高校生が作成した英作文を本研究と同様の方法で検証することで，幅広い年代における正確性の発達過程や指導方法の選択に貢献することができる。正確性は自分の意図を適切に伝える上で，非常に重要な能力である。日本人英語学習者の正確性を向上させることができる指導方法や，適切に測定できる評価方法の研究がさらに増えることが期待される。

謝辞

本研究を実施・発表する貴重な機会を与えて下さった，公益財団法人 日本英語検定協会の皆様，ならびに選考委員の先生方に心よりお礼申し上げます。特に，助言者である寺内 一先生には，有益なご助言・ご指導をしていただきました。さらに，筑波大学大学院の平井明代先生には，本研究の立案から報告書執筆に至るまで，丁寧なご指導をしていただきました。心より感謝申し上げます。最後に，本調査に協力して下さった，須田佳成さん，Angelina Kovalyovaさん，大橋伊織さんには，厚く御礼申し上げます。

参考文献（＊は引用文献）

Abe, M. (2019). Comparing errors across an L2 spoken and written error-tagged Japanese EFL learner corpus. In Gotz, S. & Mukherjee, J (Eds). *Learner Corpora and Language Teaching,* pp. 157-174.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice.* Oxford.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), pp. 267-296. https://doi.org/10.1016/S1060-3743(03)00038-9

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*(1), pp. 59-84. https://doi.org/10.1017/S0272263104026130

Evans, N. W., Hartshorn, K. J., Cox, T. L., & de Jel, T. M. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing, 24,* pp. 33-50. https://doi.org/10.1016/j.jslw.2014.02.005

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics, 36,* pp. 98-116. https://doi.org/10.1017/S0267190515000082

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), pp. 461-473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and The World, 1,* pp. 91-118.

Kato, T. (2019). Constructing measurement models of l2 linguistic complexity: A structural equation modeling approach. *JLTA Journal, 22,* pp. 23-43.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*(1), pp. 48-60. https://doi.org/10.1016/j.jslw.2007.08.003

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine - grained clausal and phrasal indices. *The Modern Language Journal,* 102(2), pp. 333-349.

Lê S., Josse J., Husson F. (2008). "FactoMineR: An R Package for Multivariate Analysis." Journal of Statistical Software, 25(1), pp. 1-18. https://doi.org/10.18637/jss.v025.i01.

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp.50-68). Taylor & Francis.

Nagata, R., Whittaker, E., & Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1210-1219).

Nation, P., & Beglar, D. (2007). A vocabulary size test. The *Language Teacher, 31*(7), pp. 9-13.

Oka, H. (2021). Reliability and optimal designs for measuring accuracy in L2 writing with a weighted clause ratio. *Annual review of English education in Japan (ARELE),* 32, pp. 49-64.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), pp. 590-601. https://doi.org/10.1093/applin/amp045

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), pp. 101-143. https://doi.org/10.1111/0023-8333.31997003

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing, 26,* 10-27. https://doi.org/10.1016/j.jslw.2014.09.003

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error - tagged EFL learner corpus. *The Modern Language Journal, 97*(S1), pp. 77-101. https://doi.org/10.1111/j.1540-4781.2012.01422.x

投野由紀夫，& 望月源．(2012).「編集距離を用いた英文自動エラータグ付与ツールの開発と評価」コーパスに基づく言語教育研究報告，No.9, pp. 71-92.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity.* University of Hawaii Press.

資料1： WCRの評価尺度（Oka, 2021より）

Level	Code	Kinds of errors
Lv.1 (0.8)	vhp	動詞の時制や分詞に関する誤り
	prep	前置詞の欠如・余剰・間違いに関する誤り
	rc	関係詞の欠如・余剰・間違いに関する誤り
	sv_agree	主語ー動詞の一致
	pro/pos	代名詞 / 所有格に関する誤り
	sg/plu	名詞の単数形 / 複数形に関する誤り
	neg	否定の欠如・余剰・間違いに関する誤り
	art	冠詞の欠如・余剰・間違いに関する誤り
	wlex	語彙の選択に関する誤り
	wwf	品詞の誤り
	m/ex_w	語彙の欠如・余剰に関する誤り
	punc	修辞法に関する誤り
	ge/inf	動名詞 / 不定詞の欠如・余剰・間違いに関する誤り
	sf	断片文に関する誤り
	sp	スペリングの誤り
	wop	語順に関する誤り
	conj	接続詞の欠如・余剰・間違いに関する誤り
	qn_agree	数詞に関する誤り
	com/sup	比較級・最上級の欠如・余剰・間違いに関する誤り
	run-o	Run-on に関する誤り
Lv.2 (0.5)	mv	be動詞の欠如に関する誤り
	mv_cop	動詞の目的語の欠如・余剰に関する誤り
	art	冠詞の欠如・余剰・間違いに関する誤り
	prep	前置詞の欠如・余剰・間違いに関する誤り
	wlex	語彙の選択に関する誤り
	wwf	品詞の誤り
	wop	語順に関する誤り
	m/ex_w	語彙の欠如・余剰に関する誤り
	ge/inf	動名詞 / 不定詞の欠如・余剰・間違いに関する
	rc	関係詞の欠如・余剰・間違いに関する誤り
	wm	法助動詞の欠如・余剰・間違いに関する誤り
Lv.3 (0.1)	nc	節の意味が読み取れない
	mv	一般動詞の欠如に関する誤り
	ms	主語の欠如に関する誤り
	mv_cop	動詞の目的語の欠如・余剰・間違い関する誤り
	rc	関係詞の欠如・余剰・間違いに関する誤り
	pv	受動態の欠如・余剰・間違いに関する誤り
	wlex	語彙の選択に関する誤り
	m/ex_w	語彙の欠如・余剰に関する誤り
	wop	語順に関する誤り