

嘉庫 嘉悦大学学術リポジトリ Kaetsu

University Academic Repository

文章によるパフォーマンス課題の自動採点の可能性

メタデータ	言語: ja 出版者: 公開日: 2025-03-31 キーワード: 作成者: 榎澤, 祐一 メールアドレス: 所属:
URL	https://kaetsu.repo.nii.ac.jp/records/2000078

研究ノート

文章によるパフォーマンス課題の自動採点の可能性

The Possibility of Automated Scoring for Performance Tasks Based on Text

榎澤 祐一*

Yuichi ENOSAWA

<要約>

近年の大学や大学院では授業での「主体的・対話的で深い学び」が要請され、かつ、学習した理論の活用という観点で、社会実装を想定したパフォーマンス課題が授業で提示されることがある。特に人文・社会科学系の科目では文章による解答を求める課題が多く、この取り組みは教育の質向上の観点からも重要である。しかし、文章解答の採点は採点者の労務管理の観点から採点時間の短縮などの効率性の向上が課題となる。そこで言語解釈能力が飛躍的に向上した生成AIであるGPT-4を用い、プロンプトにルーブリックを導入して文章によるパフォーマンス課題の自動採点を試み、その信頼性を測定した。人とGPTによる評価を比較した結果、解答の総合得点は両者で「まずまず一致している」と解釈できる結果であった。本研究は、人文・社会科学系科目の初歩的な文章によるパフォーマンス課題の自動採点の可能性を示唆する。

<キーワード>

自動採点、人文・社会科学系科目、パフォーマンス課題、ルーブリック、GPT

1 はじめに

近年の大学や大学院では「主体的・対話的で深い学び」を意識した授業が実践されつつある。大学では複数の知識の関連付けや知識を使った問題解決という観点で「深い学び」は従前から実施されてきたと考えられるが、その実践はゼミナールが中心であった。

多人数を対象とした講義での「主体的・対話的で深い学び」の実現方法のひとつに、提示された課題に対しての文章による解答と、教員によるフィードバックを通じた学習プロセスが考えられる。しかし、文章採点は選択式解答などの採点と比較して、評価者の採点基準の一貫性の問題点が指摘されており、人間の主観性や疲労 (Hussein et al., 2019) が

* 嘉悦大学経営経済学部 准教授

その要因であると示唆されている。また、労務管理の観点からは、大人数が履習する科目での解答の採点は、採点者の過度な疲労や長時間労働につながる懸念もあるだろう。学問分野別の大学教員の採点に関する数少ない研究（2021年度大学経営政策演習受講生一同、2022）によると、人文科学系教員において「コメントをつけて課題などの提出物を返却する」ことが、「教育と研究の両立の困難感」に正の影響を与えている。すなわち、課題評価作業により教育に関する時間が増大し、研究との両立に困難を感じていると解釈できる結果であった。

大学の授業の質的改善と採点業務の負荷軽減を考慮した教育的方略としてピア・ラーニング（peer learning）としての学生同士での採点が提案されることがある。この方法は学生が他者の解答を採点することにより、知識の理解が深まる利点がある。ただし、選択式問題や名詞、数値による解答を想定した問題など、解答の解釈の相違が生じにくい解答では有用であっても、文章による解答では採点する学生により採点結果に相違が生じる懸念がある。

論述式の解答を求める課題は、卒業論文やビジネス文書特有の文章技法の習得（e.g., 滝浦、2022）、論理的思考力を養成する上で有効であるため重要性が高い。そこで、採点者の負担を軽減しながら文章課題を導入することが有力な解決策のひとつになるだろう。その方略のひとつである自動採点については従来、語学科目での論述問題の採点に関する研究が進展しており一部で実用化されている。一方、人文・社会科学系科目での研究の眼目は正解の方向性が単一である際の短文解答の採点に集中している。その理由として、従来の自動採点では、特定の領域に限定された学習データのみを基盤としており、解答の文脈までの理解が困難であったことが考えられる。

しかし、2022年後半より文章の文脈把握が可能なAIとしてGPT-3に焦点が当たった後、2023年にはGPT-4が出現し、文章の解釈能力が急速に改善している。また、そのGPT-4を搭載したチャット型サービス「ChatGPT PLUS」では、同年にコードインタープリター（code interpreter）が実装された。チャットとしてのテキスト入力とともにファイルをアップロードして、そのファイルを処理できるようになったため、自動採点の実務を容易に行える条件が揃ってきている。

そこで、社会科学系大学の初年次科目を想定して、100～200字程度の短文によって解答し、正解の方向性が単一ではないパフォーマンス課題の自動採点方法を研究課題として検討した。本研究の目的は、社会科学系大学での文章によるパフォーマンス課題の解答評価における質と効率の向上である。

2 先行研究レビュー

2.1 文章によるパフォーマンス課題の評価技法

課題への回答の評価手法には、選択解答や自由記述によるテストを中心とした筆記による評価とパフォーマンス課題による評価がある（西岡、2005）。パフォーマンス課題は知識

やスキルを基に実践することを求める課題であるが、その解答方法は実演（e.g., ダンスのパフォーマンス）、実技（e.g., 調理の過程）、成果物（e.g., プログラムの成果物）など多様な形態がある。人文・社会科学系の科目では、特に文章による解答が多いだろう。

筆記による評価では、特定の法律の名称を問う設問など暗記した知識の定着度合いを評価するのに適している。一方、その知識を組み合わせて社会一般の文脈で使いこなす能力、例えば、マーケティングの授業で商品開発の観点を教示した上で新商品アイデアを文章で記述する課題を評価するには、パフォーマンス課題による評価が適している。

パフォーマンス課題の取り組みにあたっては、専門用語や理論の記憶だけでなく、現実の事例を抽象化し、理論と整合する部分を抽出したり、その分析結果を適切に文章で表現したりする能力が求められる¹⁾。そして、人文・社会科学系の大学で課される文章課題の一角を占めるのがパフォーマンス課題であり、その評価の代表的な手法がルーブリックである。ルーブリックは主に初等・中等教育機関で用いられてきたが、近年では大学をはじめとした高等教育機関でも利用されるようになってきた（沖, 2019）。

ルーブリックは「学習の到達度を測るための評価基準を表で示したもの」（山本, 2019）である。ルーブリックには単一の学習目標を評価する全体的ルーブリックと、複数の学習目標を評価する分析的ルーブリックがある（岩本, 2020）（表1）。前者でも後者でも評価目標の到達度に関する数段階の評価得点と、それに対応する評価基準の文章がある。前者の説明文では複数の評価観点が含まれることがある。さらに評価の質的側面でルーブリックを分類すると、一般的ルーブリックと課題特殊のルーブリックに分類される（岩本, 2020）。

一般的ルーブリックの場合、特定の領域であれば、どのような課題でも適用できる観点によって構成され、例えばレポートにおける「句読点の適切な運用」といった観点がある。課題特殊のルーブリックでは、特定の課題のみに適用できる観点が構成される。これらが複合的に用いられることもある。

	評価基準①	評価基準②	評価基準③
5- 優れている			
4- とてもよい			
3- よい			
2- 不十分			
1- 劣っている			
0- 課題未実施			

※評価尺度は本例では5段階評価としているが、5段階以外でも良い。評価基準①～③には、「論理的思考力」などの評価したい基準を記述する。基準の数は複数あれば3項目でなくても良い。空欄には各評価基準における評価尺度相当の達成目標を記述する（e.g., 「論述された主張を裏付ける定量的情報が記述されている」）。

表1 分析的ルーブリックのイメージ

ループリックを運用する際の課題は、これを複数名で利用した際の信頼性である。そこで、採点者間の評価差異を最小化するための方法として間主観性による担保がある。例えば、サンプル回答の評価を複数人で実施し、その採点結果を評価者間で合議する方法である（岩本, 2020）。

なお、ループリックは評価手法の側面があるだけでなく、学生が課題の自己評価に用いたり、ループリックの修正に参加したりして積極的に教育に活用する事例が存在する（寺嶋・林, 2006）。しかし、本研究は教員による課題評価に焦点を当てているため、評価面での利点を以下に述べる。

ループリックの利点の第一は、評価基準について教員と学生の間で評価の事前に共通認識をもてる点であり、これを公表と承認の原則という（西岡, 2005; 山本, 2019）。ループリックにより学生は課題に取り組む目的が明らかになり、動機付けを高める（西岡, 2005）。第二に学生は返却された評価を基に自身の回答の評価根拠を推定することが可能になる。これはパフォーマンス課題の評価において評価者の恣意性を排除し（松下, 2007）、評価の平等性（西岡, 2005）の向上にも繋がるだろう。

2.2 文章の自動採点の技法

2.2.1 日本語の人文・社会科学系の自動採点技法

自動採点の研究や実用化の多くは英文エッセイに関するものが多く、日本語の人文・社会科学系の課題の自動採点に関する研究では、その採点対象は、パフォーマンス課題の採点ではなく、入学試験などを想定し、単一の正解の方向性がある解答が中心である。

日本語の人文・社会科学系の設問への記述式解答の採点を目的とし、かつ、単一の正解の方向性が存在する解答の自動採点方法には、正解の基準を設定するものと模範解答や解答に含まれるべき単語を設定するものがある。なお、これらでは機械学習（machine learning）の手法として、ランダムフォレスト（random forest）（e.g., 石岡ら, 2016）または、ニューラルネットワーク（neural network）（e.g., 寺田ら, 2016; 竹谷ら, 2019）を利用している。

正解の基準を設定するタイプの採点方法として竹谷ら（2019）によるシステムは、解答を正誤2値判定するために解答の文字数の条件、解答に含めるべきではない単語の条件への適合性判定、正誤判定用のキーワードとの照合を行い、自動採点可能と判断した場合に自動採点する。その結果、自動採点できた解答の採点精度は高かったが、自動採点できた解答は国語、社会で60%台に留まった。この採点方法で高い精度で採点できるのは、解答の表現に制限をかけられる場合に限定される（高井ら, 2019）。

模範解答や解答に含まれるべき単語と照合した自動採点方法の研究の内、模範解答との照合による研究（石岡ら, 2016）では、大学入試の模試の社会科の記述問題を設問とし、模範解答と受験者の解答間の意味的同一性を中心に3～6点の配点で判別した。その結果、

高度な意味的判断を要さない設問において人間による採点と比較した結果、7~9割の解答が1点差に収まる成績を残している。また、正解に含まれるべき特定の単語の有無と、それらの係り受け関係を手がかりとして正誤2値判定する採点方法（寺田ら, 2016）では、大学入学希望者学力評価テストの記述式問題を大学生に解かせ、その解答を評価している。その科目のひとつに世界史の問題があり、90%近くの精度を記録している。

人文・社会科学系の文章によるパフォーマンス課題に特化した研究は無い。ただし、パフォーマンス課題と判別できる課題を含む自動採点の研究がある（竹内ら, 2017）。同研究では大学生に講義を受けてもらい、その講義テキスト、他の受講者の解答、正解例、Wikipediaを基に採点モジュールが自動採点するシステムを構築している。採点モジュールは4つあるがパフォーマンス課題と判別できる課題の採点に最も関係があるのが「理解力評価モジュール」であり、設問意図の理解を評価するものである。同モジュールでは、言語モデル n-gram により講義内容と正解例の単語の類似度を基に解答を評価する。人間の採点結果との相関係数は0.7台を記録した。

これらの研究を始めとして多くの研究では、主に機械学習を中心とした技法の洗練化による精度向上に焦点を当てている。しかし、学習データは各研究グループが収集した非公開のデータであり、記述式解答の自動採点の主要課題のひとつとして「利用できる共通の小論文データが存在しないこと（竹内ら, 2017）」が指摘されている。これはパフォーマンス課題だけの問題ではないが、単一の方向性の解答を正解としない特徴をもつパフォーマンス課題では、特に問題となる。

2.2.2 大規模言語モデルを利用した自動採点

2022年後半からプログラミング言語の入力を要さず、文字入力によるプロンプトベースで操作できる ChatGPT が注目されている。そして、操作の容易性だけでなく自動採点を行う上で重要な精度の点でも注目されており、その背景は次の通りである。

機械学習ではデータを分析する際にデータの特徴を数値化して表現した特徴量を指定する必要がある。その上で人間による採点結果と解答データをセットで用意して人工知能を訓練する。これにより機械学習では採点モデルを形成し、未知の文章の評価が可能になる。さらにそのサブカテゴリーとしての深層学習では、特徴量自体が自動生成され、文章の特徴を機械的に判別するだけでなく、文脈を捉えることが可能になる（e.g., Dong et al., 2017; Hussein et al., 2019）。

そして、深層学習が注目されるきっかけとなったのが Transformer アーキテクチャを基盤とした言語モデルである（Vaswani, et al., 2017）。Transformer アーキテクチャは、Attention（機構）という文中の重要な情報に焦点を当てて文章を解釈するための仕組みを用いており、従来手法よりもモデルの精度が向上した。これを用いた巨大なデータセットに基づく大規模言語モデルとして Google AI Language の研究者が開発した BERT と、

OpenAIが開発したGPTがある (Mizumoto & Eguchi, 2023)。特にGPTの訓練データ量をBERTと比較すると、訓練データとなる自然言語処理研究のために用いる言語データベースであるコーパスが大規模な点にGPTの特徴がある (Mizumoto & Eguchi, 2023)。

GPT-3 text-davinci-003モデルを用いて非ネイティブではない話者による英作文を採点した研究では、モデルは採点において一定の精度を維持したことが明らかになっており、モデルの微調整により精度が向上する可能性や、人間による評価と組み合わせての使用が示唆されている (Mizumoto & Eguchi, 2023)。しかし、この結果は最新の2023年11月6日にリリースされたGPT-4 Turboでの採点結果ではない。GPT-4 Turboの前身であるGPT-4のリリース時 (2023年3月14日) にOpenAIは、タスクの複雑さが一定の閾値を超えた段階でGPT-3.5との性能差が表れるとしていた (Open AI, 2023)。GPT-4 Turboでも同様にそれ以上にGPT-3.5との性能差が表れるだろう。

3 自動採点方法

3.1 採点方針

言語モデルGPT-4 Turboの利用にあたっては、チャットサービスにより指示を行うChatGPTを用いた。指示文 (以下、プロンプト) においては、採点基準として加筆・修正の容易性の観点から分析的ループリックを採用し、文章の文法項目の採点基準を定めた一般的ループリックと、設問に対する解答の妥当性を評価する課題特殊的ループリックを組み合わせた。

3.2 採点データ

筆者が担当する東京都内の私立大学の科目「マーケティング入門」(2023年12月5日に1回、12月6日に2回を異なる履修者に対して実施) で提出された国際マーケティングに関するパフォーマンス課題への解答データ (合計110件) を対象とした。解答者は全員1年生であり、殆どが日本語ネイティブである。大学でのマーケティング関連科目の履修歴はない。

解答は授業中の制限時間内にスクリーン上に投影された設問を読み取り、パソコンにより、所定のフォーム入力を介して行うものである。解答前に解答入力への支障がある際や、問題文の意図の理解や解法の導出が困難な場合は挙手の上、筆者にサポートを要求するよう呼びかけたが申し出は無かった。結果として無解答や無効解答は無かった。

解答を収集後、筆者により採点を実施し、その後、解答データをChatGPTにアップロードするとともに、ループリックを含むプロンプトを入力の上、データの採点を実施した。なお、筆者による採点時に解答に個人情報が含まれる場合は除去することとしていたが、該当する解答は存在しなかった。

3.3 採点方法

ChatGPTのプロンプトには、設問に加えて設問の意図や解答に必要な条件を記述した上でループリックを記述した（表2）。

プロンプト	プロンプトの目的
添付ファイルの内容は、設問に対する学習者の解答です。	採点データを指定する。
次の設問について、採点基準①（文法）と採点基準②（内容）に従って、それぞれ採点し、解答が入力されたセルの右隣りにそれぞれの得点を出力してください。	採点方法、得点の計算方法と出力方法を指示する。
設問 日本に進出している外国企業が、日本の文化または習慣に適応している例を、企業名または商品名を挙げて説明してください。160文字程度が目安です。	採点対象となる解答の設問を提示する。
採点基準①（文法） 3点：日本語文法に誤りがない。かつ、句読点を適切に打ってある。 2点：日本語文法に誤りがないが、句読点を打っていない。 1点：日本語文法に誤りがある。 0点：文字列の入力が無い。 採点基準②（内容） 3点：下記の解答要件のすべてを満たしている。 2点：下記の解答要件の内、3)は記述しているが、1)と2)のどちらか、または両者を記述していない。 1点：「3点」、「2点」の採点基準に当てはまらない解答。 0点：文字列の入力が無い。	採点基準のループリックを提示する。本例では評価基準を2つとしている。
解答要件 ・解答内で企業名または商品名が固有名詞で具体的に記述されている。 ・解答内で日本の風土か習慣を記述している。 ・解答内で外国企業が日本の文化または習慣に適応している事象を描写している。	

実際のプロンプトでは本表のように罫線を引かずの一続きの文章として記述している。

表2 記述したプロンプト

4 自動採点方法の評価

4.1 信頼性の評価

人が採点した結果とGPT-4 Turboが採点した結果を比較し、評価者間信頼性を算出した(表3)。具体的には文法に関する採点基準①と、解答内容に関する採点基準②それぞれの採点結果、そして両者の合計について、2次の重み付き κ 係数を算出した。 κ 係数は二者の値の観察された一致率(本例では全解答数における得点が両者一致した解答数の割合)から偶然の一致率(本例では全解答数における偶然により評価が一致したとみなされる解答数の割合)²⁾を差し引き、二者の値の実質的な一致率を算出する手法である。

そして、2次の重み付き κ 係数は順序尺度に用いる κ 係数である(Fleiss & Cohen, 1973)。重み付けのない κ 係数では名義尺度を対象に二者の値の一致・不一致のみを考慮して係数を計算するが、2次の重み付き κ 係数では二者の値が最も異なる時に0、完全一致している時に1の重みを付けるものとし、その他の場合では両者の隔たり具合(本例ではGPT-4 Turboと人の採点値の差)を考慮して重み付けする。

	κ 係数	標準誤差	p 値
採点基準①(文法)	-.012	.332	.971
採点基準②(内容)	.351	.005	<.001
合計得点	.324	.005	<.001

表3 GPTと人による自動採点結果の κ 係数

その結果、採点基準①(文法)では殆どの解答でGPT-4 Turboと人の採点値が一致し(表4)、かつ評価が3点に集中したため κ 係数は有意ではなかった。これは κ 係数の特徴として、偶然の一致率が高いと κ 係数が低くなる性質があることによるものであり、「 κ 係数のパラドックス」と呼ばれている(Byrt, Bishop & Carlin, 1993)。対象者の殆どが日本語ネイティブであるため、手書きと比較して文法的誤謬の可能性が著しく少なかったことに起因する恐れがある。また、「知能が正常であったとしても、通常の努力では文字習得が困難(宇野, 2016; p.8)」な障害である発達性読み書き障害(Developmental Dyslexia)の出現率が一定の水準にあること(宇野, 2016)を鑑みた際、ワープロ利用によって文法の正確さが向上した可能性を排除できない(Morphy & Graham, 2012)。ただし、先行研究の分析対象には日本語文章に関する研究が含まれておらず、日本語文章に関する研究は事例研究に偏っているため(河野, 2015)、日本語でのワープロ利用の効果については今後の研究が待たれる。

GPT/採点者	0点	1点	2点	3点	計
0点	0	0	0	0	0
1点	0	0	0	0	0
2点	0	0	0	1	1
3点	0	0	2	107	109
計	0	0	2	108	110

表4 採点基準①におけるGPTと人による採点結果の得点別解答数

一方、採点基準②（内容）と合計得点については、前者が.351（95%CI [.341 .361]）、後者が.324（95%CI [.314 .334]）と1%水準で有意な値を示しており、GPT-4 Turboと人の採点結果が、まずまず（Fair）一致している（Landis & Koch, 1977）と解釈できる結果となった。ただし、十分に一致していると言える結果ではなく、その原因の一端として人がGPT-4 Turboよりも高評価しやすい傾向が見て取れる（表5）。採点基準②において、採点者がGPT-4 Turboより2点高く評価している解答が13解答、1点高く評価している解答が24解答存在するのに対し、GPTが採点者より2点高く評価している解答が2解答、1点高く評価している解答が11解答に留まり、その傾向は合計得点にも引き継がれたと考えられる（表6）。

GPT/採点者	0点	1点	2点	3点	計
0点	0	0	0	0	0
1点	0	30	10	13	53
2点	0	7	19	14	40
3点	0	2	4	11	17
計	0	39	33	38	110

表5 採点基準②におけるGPTと人による採点結果の得点別解答数

GPT/採点者	3点	4点	5点	6点	計
3点	0	0	1	0	1
4点	0	30	9	13	52
5点	0	7	20	13	40
6点	0	3	3	11	17
計	0	40	33	37	110

※0点、1点、2点の解答数は省略。

表6 合計得点におけるGPTと人による採点結果の得点別解答数

4.2 考察

社会科学系の科目での文章課題、中でもパフォーマンス課題による評価において、ループリックを用いたGPT-4 Turboでの自動採点においては、高い実用性を示すまでには至らなかったが、その可能性は見て取れた。

また、GPT-4 Turbo Tの問題だけでなく、ループリックの改善も人とGPT-4 Turboの評価の一致度を上げるためには重要と言えるだろう。例えば、文章の内容評価のループリックの採点基準をより詳細にすれば、人とGPT-4 Turboとの採点の乖離を本研究よりも抑えられる可能性があるだろう。

さらに、 κ 係数を指標とする際の留意点として本研究では「採点基準①」、「採点基準②」ともに事実上の3段階評価であったが、評価段階を増やせば、二次の重みつき κ 係数の計算過程で重みづけを精緻に行うことができ、高い係数を得られた可能性もあっただろう。

5 結論

5.1 学術的示唆

本論文の学術的示唆は次の通りである。

第一に従来のループリックの課題である信頼性の問題を克服する視点を提供した点にある。すなわち、ループリックを用いた生成AIによる評価では、採点の一貫性が保持されるので採点の信頼性の課題は生じない。また、人による評価では同一人物による採点であっても一貫性の問題が生じるが、その観点においても本手法は課題を解決する。

第二に日本語の人文・社会科学系の科目における文章解答の自動採点では、単一の正解の方向性がある解答を主な対象としているが、本研究では人文・社会科学系の科目における文章によるパフォーマンス課題の採点での適用可能性を示唆した。

5.2 実務的示唆

本研究は、人文・社会科学系の科目において、利用できる解答のデータベースが無くても短文によるパフォーマンス課題の解答の自動採点をプロンプトベースの生成AIで実施できる可能性を示した。独自の学習データが無くても採点をプロンプトベースのAIで実施できることは、自動採点を容易に実施できる可能性を示唆する。さらには、文章によるパフォーマンス課題は、大学だけでなく「総合的な学習(探究)の時間」でも出題されるため、小学校・中学校・高等学校教員の採点業務にも資する可能性があるだろう。

5.3 残された課題と可能性

本研究の残された課題として、人による採点の精度の問題がある。本研究では人による採点結果を言語モデルGPT-4 Turboとの比較対象としたが、設問の出題者、採点者ともに同一人物である。設問者と採点者が異なる場合や、採点者が複数名いれば、本研究の結果

が異なる恐れがある。

第二に人工知能の利用による採点プロセスのブラックボックス化が挙げられる。特に本研究が想定した実務を前提とすると、人工知能による採点結果に対する異議が学習者などから生じた際に、その正当性を学習者に対して示すことが困難である。ただし、解答事前でのルーブリック提示などの代替策はあるだろう。

第三は妥当性の問題である。本研究では人とGPT-4 Turboの採点の近似性を基に信頼性を測定した。しかし、プロンプトなどの改良により信頼性が高まったとしても、人とGPTが採点しているものが同じであるのか、つまり妥当性の問題に帰結する。

これらの残された課題はあるものの、本研究は人文・社会科学系の科目での文章によるパフォーマンス課題における自動採点に関する議論の端緒になり得るだろう。教育の質向上や採点者の労働効率の向上のために、更なる研究と実装の試みが望まれる。

注

- 1) パフォーマンス課題による評価の他の例として、例えば書道に関する科目で、手本に則り筆を用いて文字を美しく書くことを目的とした書写能力という単一能力を評価する場合は、実技試験が用いられる(西岡, 2005)。つまり、パフォーマンス課題による評価は成果物の評価と、実演の評価(実技課題)に分類される。本論の評価対象は成果物の中でも文章を対象としているが、他の成果物(e.g., 画像、プログラム)や実技課題でも、ルーブリックに則った自動採点の可能性はある。
- 2) 偶然の一致率は、全体の解答数に対するクロス集計表(本研究では表4、表5、表6)の対角要素の期待値の合計の割合である。対角要素の期待値は、本例では人とGPTそれぞれの評価の分布に基づいて計算する。そのため、特定の評価で一致する頻度が高いと、その評価の頻度の期待値が過大になり、ひいては偶然の一致率は過大に、 κ 係数は過少になってしまう難点がある。

参考文献

- [1] 2021年度大学経営政策演習受講生一同(2022)「大学教員の教育・研究に係る両立の困難感の規定要因—学問分野ごとの特性を踏まえた分析—」『大学経営政策研究』12, pp.137-153
- [2] 石岡恒憲・亀田雅之・劉東岳(2016)「人工知能を利用した短答式記述採点支援システムの開発」『電子情報通信学会技術研究報告』116(379), pp.87-92
- [3] 岩本祐樹(2020)「ルーブリックの効果的な使い方の検討」『教育・研究』34, pp.39-52
- [4] 宇野彰(2016)「発達性読み書き障害」『高次脳機能研究』36(2), pp.8-14 doi: 10.2496/hbfr.36.170
- [5] 沖裕貴(2019)「ルーブリックとは何か」『物理教育』67(2), pp.101-104 doi: 10.20653/pesj.67.2_101
- [6] 河野俊寛(2015)「読み書き支援へのICT利用に関する研究の動向」『金沢星稜大学人間科学研究』9(1), pp.55-60
- [7] 高井浩平・竹谷謙吾・早川純平・森康久仁・須鎗弘樹(2019)「LSTMとAttentionを用いた自動採点及び採点支援の実用化に向けて」『人工知能学会第33回全国大会論文集』
- [8] 滝浦真人(2022)『学士課程教育における日本語リテラシーを考える』東北大学高度教養教育・学生支援機構 <https://www.ihe.tohoku.ac.jp/CPD/PDPonline/archive/detail.php?id=115> (閲覧日:2023年10月5日)
- [9] 竹内孔一・大野雅幸・泉仁宏太・田口雅弘・稲田佳彦・飯塚誠也・阿保達彦・上田均(2017)「小論文の自動採点に向けたオープンな基本データの構築および現段階での自動採点手法の評価」『言語処理学会 第23回年次大会 発表論文集』, pp.839-842.
- [10] 竹谷謙吾・高井浩平・清水杏奈・早川純平・森康久仁・須鎗弘樹(2019)「大規模実データにおける記

述式問題自動採点システムの検証』『言語処理学会 第25回年次大会 発表論文集』, pp.880-881

- [11] 寺嶋浩介・林朋美(2006)「ループリックの構築により自己評価を促す問題解決学習の開発」『京都大学高等教育研究』12, pp.63-71
- [12] 寺田凜太郎・久保顕大・柴田知秀・黒橋禎夫・大久保智哉(2016)「ニューラルネットワークを用いた記述式問題の自動採点」『言語処理学会 第22回年次大会 発表論文集』, pp.370-373.
- [13] 西岡加名恵(2005)「Ⅶ 学力評価のさまざまな方法 1. 学力評価の方法の分類」(pp.76-77)田中耕治(編)『よくわかる教育評価』ミネルヴァ書房
- [14] 松下佳代(2007)『パフォーマンス評価—子どもの思考と表現を評価する—』日本標準
- [15] 山本恵(2019)『ループリックに基づくレポート自動採点支援システムの研究』南山大学博士論文
- [16] Byrt, T., Bishop, J., and Carlin, J.B. (1993) “Bias, Prevalence and Kappa” *Journal of Clinical Epidemiology*, 46, pp.423-429 doi: 10.1016/0895-4356(93)90018-V
- [17] Dong, F., Zhang, Y., & Yang, J. (2017) “Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring” *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 153-162
- [18] Fleiss, L., & Cohen, J. (1973) “The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability” *Educational and Psychological Measurement*, 33, pp.613-619 doi: 10.1177/001316447303300309
- [19] Hussein, M. A., Hassan, H., & Nassef, M. (2019) “Automated Language Essay Scoring Systems: A Literature Review” *PeerJ Computer Science*, 5, e208
- [20] Landis, J. R., & Koch, G. G. (1977) “The Measurement of Observer Agreement for Categorical Data” *Biometrics*, 33(1), pp.159-174 doi: 10.2307/2529310:
- [21] Mizumoto, A., & Eguchi, M. (2023) “Exploring the Potential of Using an AI Language Model for Automated Essay Scoring” *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- [22] Morphy, P. & Graham, S. (2012) Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing*, 25, pp.641-678.
- [23] OpenAI (2023) GPT-4 <https://openai.com/research/gpt-4> (閲覧日:2023年10月5日)
- [24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017), “Attention is All You Need” *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp.6000–6010

(2024年4月30日受付、2024年6月30日再受付)