思考力評価のためのルーブリック設計と生成 AI による対話型評価システムの開発

Rubric Design for Thinking Skills Assessment and Development of Interactive Assessment System with Generative AI

山本 大翔[*1]，東本 崇仁[*2]
Hiroto YAMAMOTO[*1], Takahito TOMOTO[*2]
[*1] 千葉工業大学情報科学部
[*1]Faculty of Information and Computer Science, Chiba Institute of Technology
[*2] 千葉工業大学情報変革科学部
[*2]Faculty of Innovative Information Science, Chiba Institute of Technology
Email: s2232165WF@s.chibakoudai.jp

あらまし：被評価者の思考力を評価する手法として，レポート，プレゼン，グループワーク等がある．しかし，これらは一人の評価者によって評価されることが多く，人間の状態や環境による不安定さから主観的な評価となることが懸念される．そこで，本研究では客観的評価を行うために生成 AI を用いてルーブリックによるパフォーマンス評価を行う手法を提案する．また，被評価者の思考を最大限表現するために，生成 AI を用いた議論の場を仮想的に実現する対話型システムを開発する．
キーワード：思考力評価，対話型論証モデル，ルーブリック評価，生成 AI

1. はじめに

現在の思考力の評価はレポートによる評価が多く用いられている．しかし，レポートによる評価では被評価者の表現力不足によって自分の考えを言語化できていない可能性があり，思考力が正しく評価されていないと筆者は考えた．そこで本研究では，生成 AI との議論を用いることで被評価者の主張を対話によって深堀りし，思考を最大限表現することを目指す．

議論の場では，思考力は論証という推論形式を通じて表現されると筆者は考えた．論証とは，具体的な情報を論拠として用いることで主張を支持する推論形式である．論証では，主張に至る過程を思考する論理的思考力と主張の妥当性を反省的に思考する批判的思考力という二種類の思考力が働くことが考えられる．そこで本研究では，評価対象をこの二種類の思考力とし，それらを包括する本研究全体の思考力を「論証における思考力」と定義する．

本研究の思考力評価システムは議論の評価であるため，評価分類をパフォーマンス評価と設定し，ルーブリック評価を実践する．ルーブリック評価を行う評価物として，被評価者と生成 AI による議論履歴から被評価者の思考を構造化した思考モデルを使用する．

2. 提案手法

この章では，被評価者の思考を構造化する手法と評価に用いるルーブリックの 2 つについて紹介する．

2.1 対話型論証モデルを用いた思考の構造化

議論履歴を評価物とすると不要な情報を含む可能性や生成 AI が正しく内容を認識できない可能性があると考えた．そのため，被評価者の思考を構造化することによってこれらの問題点を解決できると考えている．構造化に使用する思考モデルとして松下[(1)]の対話型論証モデルを基に筆者が拡張したモデル（図 1）を提案する．

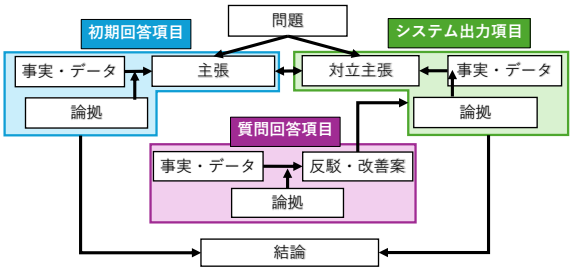

図 1 松下[(1)]の対話型論証モデルを基に筆者が拡張したモデル

2.2 本研究における思考力評価ルーブリック

この節では本研究で使用する思考力評価ルーブリックについて説明する．ルーブリックは，松下ら[(2)]のライティング・ルーブリックを基にその他思考力に関するルーブリックを参照し拡張を行った．拡張した理由は，既存のルーブリックは人間による評価のために作成されたものであり，生成 AI にそのまま認識させることが難しいと筆者が考えたためである．

作成したルーブリックは 9 観点 5 段階の構成となっており，9 観点は図 1 の構成要素を基に作成した．観点は，「背景と問題」「全体構成」「主張と結論」「統合」「仮説」の 5 観点と，初期回答項目と質問回答項目でそれぞれ分けた「事実・データ」「論拠」の 4 観点を併せた計 9 観点となっている．

3.　提案システム

提案システムは評価者用システムと被評価者提示用システムに分かれている．この章では，被評価者用提示システム(以下，ChatSystem）と評価を行うシステム（以下，EvaluateSystem）について説明する．

ChatSystem では，図 1 における質問回答項目の反駁・改善案を被評価者に入力させることで議論を行ってもらう．回答を行う生成 AI には，プロンプトとして，使用する知識や回答の方法について提示を行う．

EvaluateSystem では，生成 AI による被評価者の思考モデルのルーブリック評価を行う．評価を行う生成 AI には，2.2.で説明したルーブリックをプロンプトとして提示する．評価は，仮想的な複数人評価を行うことによって妥当な評価ができると筆者は考えたため，ルーブリック評価を複数回実行する．出力された評価結果を項目ごとで平均を求め，最終的な評価結果を算出する．

4.　予備実験

プロトタイプのルーブリックを用いたシステムによる評価の信頼性を分析するため予備実験を実施した. 実験は本大学の情報科学部の学生 18 名を対象に一人あたり最大で 2 回，国語・理科・社会の 3 科目で実施し，各科目 10 人に実施した．実験は，授業ワークを 45 分，生成 AI との議論を行うシステムの利用を 45 分，アンケートの順で行った．

実験にて得られた議論履歴から作成した思考モデルを用いてシステムの評価（評価者数 : 3）を実施した．評価結果から一般化可能性理論における G 研究を用いた分散成分の推定と一般化可能性係数の算出，D 研究を用いた各成分の増減における一般化可能性係数のシミュレーションを行った．

4.1 G 研究の結果と考察

理科では，一般化可能性係数は 0.4891 であったことから中程度の信頼性はあることが確認できた．しかし，評価として信頼性がある値（0.80）には及ばなかった．一方，国語と社会では，一般化可能性係数は 0.00 であり，信頼性がないという結果になった．信頼性が検知できなかった理由として，分散成分の推定値における被評価者の値が 0.00 であったことが作用しこのような結果になったことが考えられる．

G 研究から得られる考察として，被評価者のばらつきが全体的に検出できていないためこのような結果になったと考えられる．被評価者のばらつきが検出できなかった要因としては，ルーブリックのレベル間の表現における問題，被評価者の平均的なレベルの高さといったものが予想される．特に，レベル間の表現における問題に関しては，筆者がルーブリック作成に対する知見が少ないことが起因した可能性が高い．そのため，今後のルーブリックの作成は思考力評価の経験が豊富な評価者と議論を行い，生成 AI のルーブリック認識における多くの試行錯誤を通して改善を図っていきたい．

4.2 D 研究の結果と考察

理科における前項で得られた各分散成分の推定値を用いた D 研究を実施し，評価者数や評価観点数の増加によってどの程度一般化可能性係数が増加するのかをシミュレーションした．評価者数の増加におけるシミュレーションでは一般化可能性係数の増加は見込めないという結果になった．そのため，今回は評価観点の増加におけるシミュレーションのみを提示する（図 2）．図 2 より評価観点も増加を行っても効率的に一般化可能性係数が増加しないことが確認できる．評価観点数を 38 にすると 0.80 を超すことができるが，ルーブリックとして過剰な細分化となるので実践することはかなり難しいと考える．

D 研究から得られるシステムの問題点として，評価者や評価観点数といった量的問題ではなく，ルーブリックの内容やシステムの評価方法といった詳細な内容に問題があることが確認できた．

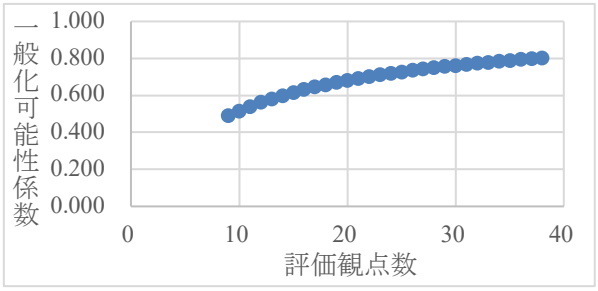

図 2 評価間点数増加における一般化可能性のシミュレーション

5.　おわりに

本研究では，生成 AI との議論履歴から被評価者の思考を対話型論証モデルとして構造化し，生成 AI によってルーブリック評価を行う思考力評価システムを提案した．また，予備実験を通してシステムの問題点を確認することができた．

今後としては，予備実験によって得られた問題点を改善し，信頼できる評価として一般化可能性係数の増加を図っていきたい．本評価実験の際には，システムの評価と実際の評価者による評価を比較し，システムの有用性を検証したいと考えている．

謝辞

本研究の一部は JSPS 科研費 JP24K00454 の助成による．

参考文献

(1) 松下佳代: “対話型論証による学びのデザイン : 学校で身につけてほしいたった一つのこと”，勁草書房（2021）
(2) 松下佳代, 小野和宏, 高橋 雄介: “レポート評価におけるルーブリックの開発とその信頼性の検討”, 大学教育学会誌, vol.35, No.1, pp.107-115 （2013）