解　説

教育システム情報学会誌　Vol. 37, No. 1　2020
pp. 8–18

テスト理論と人工知能に基づくパフォーマンス評価の新技術

宇都　雅輝*

Test Theory and Artificial Intelligence Based Technologies for Performance Assessment

Masaki Uto*

In various assessment contexts including entrance examinations, educational assessments, and personnel appraisals, performance assessment has attracted much attention to measure examinees' higher order abilities. Nevertheless, low assessment reliability and high costs of scoring are regarded as persistent difficulties hindering performance assessment. To resolve these shortcomings, item response theory models that incorporate rater and task characteristic parameters and automated essay scoring methods have been proposed recently. This paper introduces state-of-the-art topics for these technologies.

キーワード：テスト理論，人工知能，教育評価，項目反応理論，自動採点，深層学習，自然言語処理

1. はじめに

近年，さまざまな学習・評価場面において，論理的思考力や創造力，表現力などの高次な能力を測定するニーズが高まっており，そのような能力を測定する手法の一つとしてパフォーマンス評価が注目されている[(1)〜(3)]．パフォーマンス評価は，実践的・現実的な課題に対する受験者の成果物やプロセスを評価者が直接採点する評価法であり[(4)]，論述式試験やスピーキング試験，プレゼンテーション試験，実技試験，面接試験，グループディスカッションなどのさまざまな形式で活用されてきた．また，わが国では，大学入試への記述式問題の導入や英語 4 技能資格・検定試験の普及などを背景に，パフォーマンス評価のニーズは今後ますます増加すると予測できる．

他方で，パフォーマンス評価の課題として，1）人間の評価者の主観採点を伴うことによる信頼性の低下の問題[(5)〜(14)] と 2）採点コストの高さによる大規模試験実施の困難さ[(15)〜(27)] が古くから指摘されてきた．

1）の問題を解決する手法として，近年，評価者の特性を考慮して受験者の能力を推定できる数理モデルが多数提案されている[(5)(10)(12)(28)〜(33)]．これらのモデルは，情報処理技術者試験や SPI などで利用されているテスト理論の一つである項目反応理論[(34)]（Item Response Theory: IRT）に基づくモデルとして定式化されている．このような項目反応モデルは，さまざまなパフォーマンス・テストの分析や精度改善に利用されており，わが国でも医療系大学間共用試験[(35)] や英検，リクルートキャリア社などで活用されてきた．

2）の問題を解決するアプローチとしては，自動採点技術が注目されている．自動採点の研究は，主に記述・論述式試験を対象に古くからなされてきた．近年では，深層学習モデルを用いた自動採点技術が活発に研究されており，人工知能や言語処理，教育工学のトップカンファレンスである AAAI，ACL，EMNLP，AIED などで毎年新たなモデルが提案され，精度が更新され続けている[(21)〜(27)(36)〜(41)]．

以上の背景から，本稿では，パフォーマンス評価の

* 電気通信大学（University of Electro-Communications）