基盤研究(S)

※評価は非公開で行います
※一般に公開しない機微情報についても評価を受ける際に必要な内容を記入すること。

科学研究費助成事業　　事後評価報告書

令和　6 年 3月31日現在

研　究　期　間　：2019～2023

課　題　番　号　：19H05663

研　究　課　題　名　：
信頼性向上を持続するeテスティング・プラットフォームの開発

研究代表者氏名（ローマ字）：
植野 真臣（Ueno Maomi）

所属研究機関・部局・職：
電気通信大学・大学院情報理工学研究科・教授

研　究　者　番　号　：　(50262316)

交付決定額（研究期間全体）（直接経費）　　123,900,000円

研究成果の概要：
本研究では，継続的に高い測定精度と等質性のテストを実現し続けるためのeテスティング技術として，等質テスト自動生成手法，アイテムバンクマネジメント手法，等質適応型テスト手法を開発した．さらに，高品質なパフォーマンステストを継続させる手法として，評価者割り当ての最適化手法，異質評価者に頑健な項目反応理論，自然言語処理技術を用いた記述回答自動採点手法を開発した．さらに，開発したeテスティング技術を電気通信大学（入試センターと共同実施）のCBT形式の入学試験と医療系大学間共用試験のパフォーマンステストOSCEで実用化し，その知見をもとにeテスティングの運用ガイドラインを開発した．

研究成果の学術的意義や社会的意義：
本研究で開発した高い測定精度と等質性を継続できるeテスティング，パフォーマンステストの技術は，人工知能，コンピュータサイエンス，数理情報，統計学，心理学を高度に融合して実現されている．個別の要素技術が高い独創性を有することに加え，複数の領域を融合した新たな学術領域を開拓している点でも学術的に高いインパクトを有している．また，本研究で開発した技術やガイドラインに基づいて実現される高品質なeテスティングシステムは，その運用方法も含め，日本発の技術として世界の産業界に新しいマーケットを獲得できると期待できる．

１.研究組織

区分	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	宇都　雅輝 Uto　Masaki （10732571）	電気通信大学・大学院情報理工学研究科・准教授 （12612）	
研究分担者	荒木　孝二 Araki　Kouji （70167998）	東京医科歯科大学・歯学部・非常勤講師 （12602）	
研究分担者	鶴田　潤 Tsuruta　Jun （70345304）	東京医科歯科大学・統合教育機構・准教授 （12602）	
研究分担者	宮澤　芳光 Miyazawa　Yoshimitsu （70726166）	独立行政法人大学入試センター・研究開発部・准教授 （82616）	
研究分担者	大久保　智哉 Okubo　Tomoya （80512136）	独立行政法人大学入試センター・研究開発部・准教授 （82616）	削除：2020年1月17日
研究分担者	繁桝　算男 Shigemasu　Kazuo （90091701）	慶應義塾大学・社会学研究科（三田）・訪問教授 （32612）	

２．研究開始当初の背景

近年，異なるテストを受けたにもかかわらず同一尺度上で評価できる e テスティングが注目されている（植野 2009）．e テスティングは，国際標準 ISO 規格（ISO/IEC 2007）で規定されて以来，我が国最大の国家試験である情報処理技術者試験や医療系大学間共用試験（全国の医学部・歯学部の学生が合格しなければならない準国家試験），リクルート社人材測定テスト SPI，人事院公務員試験などでも実施されるようになった．また，文部科学省は「大学入学共通テスト実施方針」（文部科学省 2017）で e テスティング形式の入試について議論を開始している．e テスティングが国際標準を満たすための重要な要件は，異なる項目で構成されているにも関わらず等質なテストを多数生成しなければならないことである．等質テスト生成とは，あらかじめ蓄積されたアイテムバンクから等質性の条件を満たすテスト群を可能な限り多く生成することである．申請者のグループはこの手法で IEEE Transaction on Learning Technologies, Computer Society やトップカンファレンス AIED に多数の論文を掲載し，関連成果は複数の学会から賞を授与されている．しかし，実際の大規模テストを長期間運用することにより，年度の経過に応じて誤差が徐々に増加してくる新たな問題が生じている．一般的な等質テスト生成法では若干の問題項目を等質テスト間で重複して良いように設計されている（申請者のグループでは，一つのテストの項目の 5%が他テストの項目と重複して良いとしている）．実際の測定誤差が経年的に増加する原因は，自動テスト構成システムが特定の良質の問題項目を集中的に選択してしまうことで露出率が大きくなり，受検対策などにより項目の特性が経年的に変化し，結果として測定誤差の劣化が急速に進んだことにあると考えられる．そこで，この分析を踏まえたうえで，本研究課題の第一の学術的問いは，「継続的に高精度の e テスティングを実施できる手法はどのようなものか？」である．

他方で近年，筆記試験や実技試験，面接試験のような評価者を伴うパフォーマンステストの信頼性向上のニーズが高まっている．例えば，医療系大学間共用試験では OSCE と呼ばれる実技試験，英検では Speaking や Writing の試験などが知られている．これらの人間の評価者を伴うテストでは，評価の信頼性が評価者の特性（厳しさや一貫性，識別性など）に依存することが問題である．申請者らは，評価者の特性をパラメータ化し，それらを数理モデルに組み込んでより信頼性の高い評価を実現するための項目反応理論を開発してきた．この技術によって，スコア測定に対する評価者特性の影響を軽減することに成功したが，それらの特性がスコアの測定誤差に与える影響は依然として大きい．そこで，本研究課題の第二の学術的問いは，「継続的に高精度のパフォーマンステストを実施できる手法はどのようなものか？」である．

３．研究の目的

本研究の目的は前述の二つの学術的問いに対応し，以下の通りである．

目的（1）：継続的に高精度のテストを生成し続けるための技術・運営法を提供し，実際の大規模テストの運営に適用して実証する．

目的（2）：パフォーマンステストにおけるスコアの測定誤差を等質かつ継続的に減少させていく手法の開発を行い，実際のパフォーマンステストに適用して実証する．

目的（1）を実現するために以下を実行する．1．等質テストの生成数をよりダイナミックに向上させ，その上でアイテムバンクからの項目の露出を一様にするようにテスト構成できるアルゴリズムを開発する．2．劣化している項目を検出する手法を開発し，検出された項目をデータベースから削除する．また，劣化する項目数および削除すべき項目数を事前に予測する手法を開発し，必要な新問題項目数を予測して作成する．さらに，新規項目が追加された場合のアイテムバンクからの等質テスト生成の効率的なアルゴリズムを開発する．3．受検者の回答履歴に基づき，適応的に問題項目を出題すれば測定精度を落とさず出題項目数を減少させることができ，項目露出を減少させることができる．しかし，適応型テストは能力値が近い受検者に重複した項目を提示することが多く，項目露出の偏りを増長させてしまう可能性が高い．そこで，各受検者にできる限り精度を保ち，異なる項目を提示し，問題の項目露出を一様分布に抑える適応型テストを開発する．

目的（2）に対しては以下を実行する．4．パフォーマンステストでは，通常の問題項目を用いた場合とは異なり，評価者セッティングのコントロールが難しいため，測定誤差の等質性と高精度性は保証されてこなかった．パフォーマンステストの等質性と高精度性を保証できるように，e テスティングの等質テスト構成法を応用し，評価者と受検者の組で測定誤差が最小になるように最適化するアルゴリズムを開発する．さらに継続的に測定誤差を減少できるように異質な評価者の検出およびそれらの評価者へのトレーニング手法の開発手法を提案する．また，評価者の負担を減少させるように受検者の実技をビデオ動画等でサーバに共有し，測定精度を向上させるように適応的に評価者を選択するアダプティブなパフォーマンス評価システムを提案する．特に近年ニーズが高まっている筆記試験については，評価者パラメータを持つ項目反応理論を用いて，自然言語処理技術を用いた現存する複数の自動採点機を人間評価者に混在して学習さ

せ，複数の自動採点機で人間の評価を近似する新しい自動採点システムを開発する．
さらに，5．開発されたプラットフォームを現実の複数のテスト場面に適用し，実証実験を行うとともに，その知見に基づいてｅテスティングの運用ガイドラインを作成する．

４．研究の方法

［ ① 研究方法 ］

1．等質テスト自動生成システム
長期的に高精度を継続させるためには従来の手法の制約を除いてより多くの等質テストを生成できる手法を開発しなければならない．これまで申請者らはテストの自動生成を最大クリーク問題として解くアルゴリズムでテスト構成数を従来手法の数倍から数十倍へと向上させてきた．本研究では，このアルゴリズムを改良することで，テスト生成数を向上させるとともに，項目露出を一様とする等質テスト自動構成システムを開発する．

2．高精度を持続するためのアイテムバンク・マネジメント手法の開発
アイテムバンク中の劣化項目の検出手法を提案する．具体的には，年度ごとに新しく得られたデータについて項目パラメータを推定し，昨年より劣化している項目を削除する．これらの劣化項目数を信頼性評価で用いられるワイブール分布を改良し，時間変数と初期項目パラメータ，露出数を変数として劣化確率を予測するモデルを提案する．また，従来の等質テスト構成手法では，新規項目が追加されるごとにアイテムバンク中のすべての項目を用いて等質テストを構成しなければならず，計算コストが膨大であった．そこで，新規項目追加時のより効率的な等質テスト構成アルゴリズムを提案する．

3．項目露出を制御する等質適応型テスト
受検者のテストへの反応結果から逐次的に能力推定を行い，情報量を最大にする項目を出題する適応型テストは，テストの測定精度を損なわずに出題項目数を減らすことができる．しかし，テストの継続的運用に際し，情報量が高い良質な項目が高頻度で出題されてしまい，受検者への項目内容の暴露に繋がり，テストの信頼性の低下要因となりうる．また，同じ能力の受検者が繰り返しテストを受けたときに同じ項目が出題されてしまう．他方で近年，大量の等質テストを生成できる手法が開発されてきた．このような等質テスト生成技術に各項目の露出率をできる限り一様にする制約条件を付加し，事前に等質テストを大量に生成した後で，その等質テストから適応的に項目を出題すれば，項目露出を制御する適応型テストが実現できる．具体的には，まず，アイテムバンクの項目露出をできる限り一様とする等質適応型テストを提案する．

4．パフォーマンステストにおける信頼性向上手法の開発
筆記試験や実技試験のようなパフォーマンステストでは評価者による採点が必要になる．しかし，この場合，評価者の特性差により，評価にバイアスが生じることが問題となる．この問題を解決するために，申請者らは，評価者の特性を考慮してスコアを推定できる項目反応理論を開発し，世界最高の測定精度を達成してきた．本研究では本技術のパフォーマンステストへの継続的適用において，等質かつ高精度な測定精度を維持するために以下の研究を行う．1）パフォーマンステストの等質性と高精度性を保証できるように，ｅテスティングの等質テスト構成法を応用し，評価者と受検者の組で測定誤差が最小になるように最適化するアルゴリズムを開発する．2）継続的に測定誤差を減少できるように異質な評価者の検出およびそれらの評価者へのトレーニング手法の開発手法を提案する．3）評価者の負担を減少させるように受検者の実技をビデオ動画等でサーバに共有し，測定精度を向上させるように適応的に評価者を選択するアダプティブなパフォーマンス評価システムを提案する．4）特に近年ニーズが高まっている筆記試験について，評価者パラメータを持つ項目反応理論において，自然言語処理を用いた現存する複数の自動採点機を人間評価者に混在して学習させ，複数の自動採点機で人間の評価を近似する新しい自動採点システムを開発する．

5．実証実験とガイドラインの開発
本研究で開発するプラットフォームを実践的に実証評価する．医療系大学共用試験では申請者らの開発した等質テスト自動生成システムが稼働しているが，本研究で開発されるプラットフォームを新たに適用し，実践的に評価する．また，医療系大学間共用試験では，臨床の実技テストとして OSCE という実技形式のパフォーマンステストが実施されている．本研究では，東京医科歯科大学における OSCE を対象に，本研究で開発するパフォーマンステストにおけるプラットフォームの実証実験を行う．さらに，大学入試センターでは，大学入試におけるｅテスティングの実用化に向けた実証実験を開始する．
また，高い測定精度のｅテスティングを実現するには，テストを実施する組織の運用方法も適

切でないといけない．そこでｅテスティングの運用をまとめ，運用ガイドラインを作成する．

［ ② 研究を遂行する上で生じた問題点及びその解決方法 ］

研究期間の前半は，新型コロナウィルス感染症拡大によって一部の研究活動において制限が生じたが，当該期間は基礎技術の開発を中心的に行う計画となっていたため全体に大きな遅延は生じなかった．ただし，東京医科歯科大学での被験者実験については，新型コロナウィルス感染症対策による活動制限により制限が生じたが，遠隔でも実験を実施できるように，Web 遠隔システムを導入した診療実技に係る新たな録画システムを開発し，この問題に対処した．

［ ③ 当初に予定していた研究経費の使用計画を変更して行った研究計画・研究方法 ］

上述の通り，東京医科歯科大学での実証実験について，新型コロナウィルス感染症対策による活動制限により対面での実験に制限が生じた．そこで，遠隔でも実験が実施できるように，当初予定に加えて Web 遠隔システムを導入した診療実技に係る新たな録画システムの開発を行った．

［ ④ 中間評価で受けた指摘事項に対する対応状況 ］

中間評価結果では，「順調に研究が進展しており，期待どおりの結果が見込まれる」と評価いただいた．個別の技術開発については，十分な成果が出ていると評価いただいており，今後は実証実験とガイドラインの開発についての成果を期待するとコメントいただいた．その後，実証実験に関しては，大学入試センターと協力して電気通信大学の入学試験において，ｅテスティングの実用化を行った．その成果は様々なシンポジウムやメディアで広く公開した．また，医療系大学間共用試験での実用も継続して行なっており，その成果は全国向けの説明会や年間報告書などで毎年報告している．さらに，これらの実践での知見を踏まえてｅテスティングの運用ガイドラインを開発し，研究代表者が管理する本科研費のウェブサイトで公開した．

５．研究成果

［ ① 本研究課題による研究成果 ］

1. 最大クリークと並列整数計画法によるハイブリッド等質テスト自動生成法

Ishii, Songmuang & Ueno（2014）は等質テスト生成問題を最大クリーク問題として定式化し，さらに大きなクリーク探索を可能とするために近似アルゴリズムを提案し，当時画期的な 10 万の等質テスト生成に成功している．しかし，最大クリークアルゴリズムは最先端のものを用いても時間計算量は低いが空間計算量が大きく申請者らの計算機環境でも 10 万程度の等質テスト生成が限界であった．その問題を解決するために石井・赤倉・植野（2017）は，より空間計算量の低い整数計画法を用いて逐次的に最大クリークを探索する手法を提案しており，約 2 倍の 194575 個の等質テスト構成に成功した．この手法は時間計算量が Ishii, Songmuang & Ueno（2014）より大きいという問題がある．本研究で提案するアルゴリズムは，コンピュータのメモリが許す限り，空間計算量は大きいが時間計算量の小さい最大クリーク近似アルゴリズムを用いてクリークを探索し，コンピュータのメモリが限界になると時間計算量は大きいが空間計算量が小さい整数計画法にスイッチするハイブリッド法である．後半の整数計画法は時間計算量が大きいので，整数計画法による S 個ずつの並列計算で現在の最大クリークサイズを増加できるアルゴリズムを提案していることが特徴である．この手法により，438950 個とこれまでに 2 倍以上の等質テストを構成できた．本研究の一部は電子情報通信学会論文誌（渕本・植野, 2020）に掲載され，さらに改良した完全な最終アルゴリズムは IEEE Transaction on Learning Technologies（Fuchimoto, Ishii & Ueno, 2022）に掲載されている．しかし，ハイブリッド法を用いても約 45 万のテスト生成に 1 ヶ月も必要である．本研究ではより大規模なｅテスティングでの運用を想定した Zero-suppressed Binary Decision Diasgrams（ZDD）を用いた新しい自動並行テスト構成手法を提案した[3,6]．ここで，ZDD とは二分決定木の圧縮表現で，組合せ集合を効率良く列挙・集合演算できる．本研究では，各節点は各問題項目と対応し，各節点はその問題を出題するか否かで二分する．さらに，受検者の測定誤差とテストの長さが等価な節点を共有する．これにより，根節点から 1（1-終端節点）までの全ての経路がそれぞれ同一の測定精度を持つテストとなる．この手法により，従来手法を大きく上回る約 150 万のテストを 1 日以内に生成できた．本研究の一部は人工知能学会論文誌（渕本・湊・植野，2022）に掲載され，さらに改良した完全な最終アルゴリズムは IEEE Access（Fuchimoto, Minato & Ueno, 2023）に掲載されている．

本研究は研究計画の「1. 等質テスト自動生成システム」に対応する．また，逐次的に新規項目が追加された場合に整数計画法の並列アルゴリズムでクリークを更新できるので「2. 高精度を持続するためのアイテムバンク・マネジメント手法の開発」にも対応する．

[参考文献] Takatoshi Ishii, Pokpong Songmuang, Maomi Ueno（2014）Maximum clique algorithm and its approximation for uniform test form assembly. IEEE Transactions on

Learning Technologies, Vol.7, no.1, pp.83-95.

2. Deep-IRT の開発

e テスティングはテスト理論における項目反応理論 Item Response Theory（IRT）を用いて受検者スコアの誤差を予測する．アイテムバンクを構築する際にあらかじめテストデータを採取し，項目反応理論を用いてあらかじめパラメータ推定しておく必要がある．この際，テストのデザインに等化もしくは Calibration を行う必要がある．これらの精度が低く，新規項目を追加する場合の大きな支障になっている．正規分布からの受検者スコアのランダムサンプリングを仮定している項目反応理論では現実のデータに合わないことも多い．申請者らはランダムサンプリングを仮定しない Deep Learning を用いた項目反応理論を提案している．受検者深層ネットワークと項目深層ネットワークから項目への反応を予測するモデルで等化などの操作を必要とせず，受検者のランダムサンプリングが難しい場合にも高い予測精度が示された．また，学習者の能力成長を考慮しながら学習者の多次元スキルに対する能力推定，項目への反応予測を行う Deep-IRT の開発も行ってきた．提案手法は多次元の能力値を表す潜在変数をもち，時点ごとに過去の学習データの忘却と新たな反応データを用いた更新を行うことで高精度な反応予測と能力推定を行うことができる．本研究では提案手法が最先端の既存手法の予測精度を上回る成果を達成した．本研究に関する成果は IEEE Transaction on Learning Technologies に 1 件（Tsutsumi et. al, 2024）掲載され，電子情報通信学会論文誌に 5 件（堤・西尾・植野,2024; 堤・郭・植野,2023; 堤・木下・植野, 2021; 木下・植野, 2020; 堤・木下・植野, 2020）に掲載されている．また，トップ国際会議である AIED・EDM に 3 件（Tsutsumi et. al 2021, 2022, 2024）採録されている．さらに人工知能学会全国大会（堤ら,2019）と教育システム情報学会全国大会（堤・植野, 2021）では大会優秀賞を受賞し，堤・郭・植野（2023）は電子情報通信学会論文誌 2023 年度論文賞を受賞した．本研究は研究計画「2. 高精度を持続するためのアイテムバンク・マネジメント手法の開発」に対応する．

3. 項目露出を制御する整数計画法を用いた等質テスト構成と適応型テストへの応用

Ishii & Ueno（2015）は Ishii, Songmuang & Ueno（2014）の等質テスト構成法を用いて等質テスト集合（クリーク集合）をすべて保存し，その中から最も露出率（＝露出数の最大値／テスト構成数）が小さい等質テストを選択する手法を提案している．しかし，この手法では項目露出の分布の偏りは大きく軽減されないという問題もある．本研究では，整数計画法を用いて現在のクリーク（等質テスト）に隣接するノード（テスト）を逐次探索して追加する過程で，この時点での露出数が上位 N 位の項目全てを逐次的に候補から削除して等質テストを追加する手法を提案した．この手法では，逐次的に露出数が上位 N 位の項目が変化していき，結果的に等質テストでの項目露出分布が一様に近づいていく．実験の結果，本手法は等質テスト構成数を減少させることなく露出率を減少できることがわかった．本研究に関する成果は電子情報通信学会論文誌（植野・渕本・植野，2022）に掲載されている．さらに，本研究では露出数の偏りを防ぐために，露出数を所与としてロジスティック関数による二種類のペナルティ項を 5.①1 のハイブリッド法における整数計画法の目的関数に追加することを提案した．一つ目はこのロジスティック関数を用いた決定論的ペナルティ項である．この決定論的ペナルティ項は露出数に応じた負の重みを常に各項目の決定変数に与える．2 つ目はロジスティック関数を用いた確率論的ペナルティ項である．この確率論的ペナルティ項は数理計画法の Big-M 法に基づいて，露出数に応じた確率により，大きな負の重みを各項目の決定変数に与える．この結果，従来手法と比較して，テスト構成数を減少させることなく，露出数の偏りを抑制できた．本研究に関する成果は統計数理（渕本・植野，2024）に掲載予定である．

従来の適応型テストは良質の項目の項目露出数が偏って大きくなってしまい結果として劣化を早めてしまう欠点があった．そこで項目露出を制御する等質テスト生成技術を用いて等質なアイテムバンクを複数生成し，各等質アイテムバンクより適応的に項目を出題すれば，項目露出を制御する適応型テストを開発した．さらに等質アイテムバンクより能力推定値が収束した後，全体のアイテムバンクから項目を出題すると項目露出を一様にできるだけでなく推定精度も従来の適応型テストから大きく劣化しないことを示した．これらの成果はトップカンファレンス AIED に Ueno and Miyazawa（2019, 2021）, Kishida, Fuchimoto, Miyazawa, Ueno（2023）で発表されている．また，適応型テストに適用する場合，等質テストを考慮する項目選択に時間を要してしまう問題がある．そこで受検者の正誤反応によって出題される項目をあらかじめ計算し，二分木を作る手法を開発している．木が大きくなると使用メモリがオーバーしてしまうために二分木を圧縮する手法も開発した．この手法により大規模なアイテムバンクから等質適応テストを実施することができるようになる．本研究の成果の一部は，人工知能学会全国大会国際セッションで Excellence Award を受賞している．本研究は研究計画「3.項目露出を制御する等質適応型テスト」に対応する．

[参考文献] Takatoshi Ishii, Maomi Ueno (2015) Clique algorithm to minimize item exposure for uniform test forms assembly. International Conference on Artificial Intelligence in

Education, pp.638-641.

4. パフォーマンステストにおける信頼性向上手法の開発

筆記試験や実技試験のようなパフォーマンステストでは評価者による採点が必要になるが，この場合，評価者の特性差により，評価にバイアスが生じることが問題となる．この問題を解決するために，申請者らは，評価者の特性を考慮してスコアを推定できる項目反応理論を開発し，世界最高の測定精度を達成してきた（Uto & Ueno，2016）．本研究課題では，本技術のパフォーマンステストへの継続的適用において，等質かつ高精度な測定精度を維持するために以下の研究を行った．下記の研究は，上述した研究計画の「4.パフォーマンステストにおける信頼性向上手法の開発」に対応している．

4.1.評価者と受検者の組の最適化アルゴリズムの開発

パフォーマンステストの精度は，評価者と受検者の組み合わせに依存すると考えられる．例えば，能力の低い受検者に厳しい評価者ばかりを割り当てた場合，得られるスコアが最低点に偏ってしまい，受検者の能力評価を適切に行うことが困難となる．そこで，本研究では，e テスティングのアプローチを用いて，評価者と受検者の組み合わせを最適化するアルゴリズムを開発した．具体的には，スコアの測定誤差を受検者全体で最小化するように評価者を割り当てる整数計画問題として定式化し，実データによりその有効性を示した．本研究の成果は IEEE Transaction on Learning Technologies（Uto, Thien & Ueno, 2020）に掲載された．

4.2.異質評価者に頑健な項目反応モデルの開発

等質で高精度な評価を継続するためには，評価者の質の維持が重要となる．評価者の質向上のためには，異質性の強い評価者を同定し，必要に応じて適切なトレーニングを与える必要がある．これを実現するために，本研究では，評価者の多様な特性を推定できる項目反応モデルを開発した．また，このモデルを利用することで，異質性の高い評価者を同定し，その特性を客観的に分析できることを確認した．本研究の成果は，Behaviormetrika，Springer（Uto & Ueno, 2020）に掲載された．

さらに，本技術の拡張として，ルーブリックを用いた評価のための項目反応モデルと評価者特性の時間変動を考慮した時系列型項目反応モデルの開発も行なった．具体的には，評価者や課題の特性に加えて評価項目の特性も考慮した能力測定を可能とする 4 層型の拡張モデルと，ルーブリックの背後に仮定される多次元の能力尺度を評価できる多次元型の拡張モデル，さらに，各評価者が多数の受検者を長時間かけて採点するような場合に評価の厳しさが時間とともに変化する「評価者特性ドリフト」と呼ばれる現象を捉えることができる時系列型の拡張モデルを開発した．これらの成果は，Behavior Research Methods（IF=5.95）を含む国内外の複数の査読付き論文誌に計 9 件が掲載され，トップ国際会議の AIED では論文賞にノミネートされた．

さらに，医療系大学間共用試験の実技試験 OSCE の状況に特化した項目反応モデルの開発を行い，東京医科歯科大学で収集した実際の OSCE データに適用する実験を行なった．本研究の成果は，現在，医学系の論文誌に投稿中である．

4.3.パフォーマンステストのためのリンケージ・デザインの設計と評価

パフォーマンステストのための項目反応理論を利用するためには，事前に各評価者の特性値を推定する必要がある．一方で，長期的に評価を運用する場合，評価者集団には入れ替わりが生じると考えられる．このような場合，新たに参入する評価者の特性値と既存の評価者集団の特性値の尺度を合致させる「リンケージ（Linkage）」と呼ばれる作業が必要となる．リンケージのためには適切なテストデザインが必要となるが，高精度なリンケージを達成するデザインの条件についてはこれまで明らかにされてこなかった．そこで本研究では，大規模なシミュレーション実験によって，項目反応理論に基づくパフォーマンステストのリンケージ精度を様々な条件下で評価した．本研究の成果は，現実場面のパフォーマンステストデザインの設計において有益なデータを提供する．本研究の成果は Behavior Research Methods, Springer に掲載された．

4.4.項目反応理論と深層学習を利用した高精度な小論文自動採点技術の開発

筆記試験においては，評価データだけでなく解答文のテキスト情報もスコアリングに利用できる．そこで本研究では，パフォーマンステストのための項目反応理論によるスコアリングに，自然言語処理技術に基づく最先端の自動採点機による予測スコアを統合する手法の開発を行なった．さらに，自動採点機そのものの高精度化や，項目反応理論と自動採点技術を融合した新たな技術の開発も行なった．具体的には，次の研究を行なった．1）評価者特性を考慮した項目反応理論を用いて訓練データ内のバイアスの影響を取り除くことで，頑健な自動採点モデルを構築できるフレームワークの開発．2）現在主流の深層学習ベースの自動採点機に，古くから利用されてきた特徴量ベースの自動採点機を統合することで得点予測の精度を向上させる技術の開発．3）論述構造解析と呼ばれる先端技術を用いて文章の論理構造を推定し，その情報を明示的

に活用して自動採点できる技術の開発，4）客観式テストへの回答履歴から推定される受検者の能力値を補助情報として活用することで，記述式課題に対する自動採点の精度を改善できる技術の開発．5）単一の総合得点に加えて複数の細目別得点も予測できる複数観点自動採点技術として，深層学習に基づく先端モデルに多次元項目反応理論を融合することで，高精度を維持しつつ高い解釈性を実現できる手法の開発．6）評価者特性を考慮した項目反応理論に基づいて多様な特徴の自動採点モデルを統合（アンサンブル）することで，高精度な自動採点を達成する手法の開発．7）自動採点技術を応用することで，通常の手法ではリンケージ不可能なパフォーマンステストのデザインでもリンケージできるようにする手法の開発．

以上の成果は，IEEE Transactions on Learning Technologies に 2 件，電子情報通信学会論文誌に 3 件を含む計 9 件の査読付き論文誌に掲載され，トップ国際会議である AIED と COLING などの累計 7 件の国際会議にも採択された．さらに，教育システム情報学会論文誌に掲載された内田・宇都（2021）は論文賞を受賞し，Uto & Okano（2020）は AIED において Best Paper Runner-up を受賞した．加えて，関連する成果は，人工知能学会や電子情報通信学会などを含む国内の様々な学会で累計 12 件の賞を受賞した．

5. 実証実験とガイドライン開発

本研究成果を用いて大学入試センターと協力して電気通信大学は新入生の基礎学力調査のためのアイテムバンクを開発し，平行テストを生成して実際に 2023 年度に実施している．このアイテムバンクは文部科学省委託事業として 2024 年度電気通信大学の総合型選抜，学校型推薦でも CBT に用いられることになっている．これらのプロセスを解析し，ガイドラインを執筆し，以下の URL に公開している．

【国際標準を満たす CBT 運用・実施のためのガイドライン】
http://www.ai.lab.uec.ac.jp/wp-content/uploads/2024/04/CBTguide.pdf

医療系大学間共用試験では，臨床の実技テストとして OSCE と呼ばれるパフォーマンステストが実施されている．本研究では，各年度に全国の医学部歯学部を対象に実施された OSCE の実データの一部を利用して，評価者特性を考慮した項目反応理論の実証実験を行った．結果として，項目反応理論の利用が，スコア測定誤差の低減に寄与することや評価者トレーニングなどへの応用に有益であることが示された．本成果は，医学教育学会全国大会（2019, 2020）および毎年実施される試験信頼性向上部会講演会（2019, 2020, 2021, 2022, 2023）で発表した．

また，東京医科歯科大学においても実技試験 OSCE を対象に，評価者特性を考慮した項目反応理論に関する実証実験を行なった．なお，研究期間前半は，新型コロナウィルス感染症対策により対面での被験者実験が困難であったため，シミュレーターにおける診療評価のための Web 会議システムと Web 学習支援システムを基盤とした評価システムを構築した．共用試験歯学系 OSCE テーマにも含まれる，医療安全・感染対策・各診療手技の評価について，Web 会議システムでの録画動画を用い Web 学習支援システム上で複数評価者が遠隔評価できることが確認された．新型コロナによる制限が緩和された 2021 年以降には，このシステムを利用して，東京医科歯科大学歯学科 6 年を対象に，OSCE 形式での臨床実習後医療面接試験のデータを収集する実践を行なった．得られたデータに，4. で開発してきたパフォーマンス評価のための項目反応理論を適用し，OSCE における評価者や課題，評価項目の特性分析などを通して，試験の信頼性や改善点を調査した．本研究の成果は，現在医学系の査読付き論文誌に投稿中である．

6. 科研費シンポジウムの開催（期間中 2 回実施）

2021 年 1 月 29 日にオンラインで本科研費成果に関する第一回公開シンポジウムを無料で開催した．ウェブサイト上（http://www.ai.lab.uec.ac.jp/kakenhi/）では報告論文集も公開した．参加者数は 231 名であった．

さらに 2023 年 12 月 27 日にはオンラインで本科研費成果に関する最終報告のための公開シンポジウムを無料で開催した．ウェブサイト上（http://www.ai.lab.uec.ac.jp/kakens2023/）では報告論文集も公開した．参加者数は 106 名であった．実施後には，「講演会全体の印象」と「講演内容の有用性」について 5 件法のアンケートを行い，30 件の回答があった．「講演会全体の印象」では「非常に良かった」が 27 件，「良かった」が 3 件であった．「講演内容の有用性」については「非常に有用だった」が 26 件，「有用だった」が 4 件であった．自由記述回答における代表的なコメントは次のとおりであり，高い評価と今後への期待が示される内容となった．

- e テスティングの最新の動向を知ることができて有意義でした．
- IRT や生成 AI の活用に新鮮さを感じました．今後の研究の発展に期待しています．
- 定期的にシンポジウムを拝聴したいと思います．
- どの発表も興味深く拝聴しました．若い研究者の精力的な活動が素晴らしいと思います．
- 数理モデルの詳細や具体的な事例をもっと提示してほしいと感じました．

［ ② 当初に予見していなかった新たな展開等によって得られた研究成果 ］

該当なし

６．主な発表論文等

〔雑誌論文〕（計　65　件）

[1] ＊渕本壱真，植野真臣（2024）項目露出ペナルティを用いた整数計画法により自動並行テスト構成．統計数理（再録決定済み）．［査読あり］

[2] ○＊Emiko Tsutsumi, Yiming Guo, Ryo Kinoshita, Maomi Ueno (2024) Deep Knowledge Tracing Incorporating a Hypernetwork With Independent Student and Item Networks. IEEE Transactions on Learning Technologies, Vol.17, pp.951-965.［査読あり］

[3] ＊Kazuma Fuchimoto, Shin-ichi Minato, Maomi Ueno (2023) Automated Parallel Test Forms Assembly using Zero-suppressed Binary Decision Diagrams. IEEE Access, Vol.11, pp.112804-112813.［査読あり］

[4] ○Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, Maomi Ueno (2023) Integration of Prediction Scores From Various Automated Essay Scoring Models Using Item Response Theory. IEEE Transactions on Learning Technologies, Vol.16, No.6, pp.983-1000.［査読あり］

[5] *Masaki Uto (2023) A Bayesian Many-Facet Rasch Model with Markov Modeling for Rater Severity Drift. Behavior Research Methods, Vol.55, pp.3910-3928.［査読あり］

[6] *柴田拓海，宇都雅輝（2023）多次元項目反応理論と深層学習を用いた複数観点同時自動採点手法．電子情報通信学会論文誌 D．Vol.J106，No.01，pp.47-56.［査読あり］

[7] *宮澤芳光，植野真臣（2022）高精度能力推定を保証する 2 段階等質適応型テスト．電子情報通信学会論文誌 D．Vol.J106，No.01，pp.34-46.［査読あり］

[8] *Akitaka Hattori, Ken-Ichi Tonami, Jun Tsuruta, Masayuki Hideshima, Yasuyuki Kimura, Hiroshi Nitta, Kouji Araki (2022) Effect df haptic 3D Virtual reality dental training simulator on assessment of tooth preparation. Journal of Dental Sciences, Vol.17, No.1, pp.514-520.［査読あり］

[9] ○＊Kazuma Fuchimoto, Takatoshi Ishii, Maomi Ueno (2022) Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly. IEEE Transactions on Learning Technologies, Vol.15, No.2, pp.252-264.［査読あり］

[10] ＊渕本壱真，湊真一，植野真臣（2022）Zero-suppressed Binary Decision Diagrams を用いた自動テスト構成．人工知能学会論文誌，Vol.37，No.5，pp.1-11.［査読あり］

[11] 植野晶・渕本壱真・＊植野真臣（2022）項目露出を考慮した整数計画法による等質テスト構成．電子情報通信学会論文誌 D，Vol.J105，No.8，pp.485-498.［査読あり］

[12] *宇都雅輝（2022）ルーブリックを用いたパフォーマンス評価のための多次元 4 相型項目反応モデル．電子情報通信学会論文誌 D，Vol.J105，No.07，pp.457-469.［査読あり］

[13] *Maomi Ueno, Kazuma Fuchimoto, Emiko Tsutsumi (2021) E-testing from artificial intelligence approach. Behaviormetrika, Vol.48, No.2, pp.409-424.［査読あり］

[14] 岡野将士，*宇都雅輝（2021）評価者バイアスの影響を考慮した深層学習自動採点手法．電子情報通信学会論文誌 D．Vol.J104，No.08，pp.650-662.［査読あり］

[15] *青見樹，堤瑛美子，宇都雅輝，植野真臣（2021）項目反応理論による小論文自動採点機のモデル平均．電子情報通信学会論文誌 D，Vol.J104，No.11，pp.784-795.［査読あり］

[16] ○*Masaki Uto, Masashi Okano (2021) Learning Automated Essay Scoring Models Using Item Response Theory-Based Scores to Decrease Effects of Rater Biases. IEEE Transactions on Learning Technologies, Vol.14, No.6, pp.763-776.［査読あり］

[17] *Masaki Uto (2021) A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. Behaviormetrika, Vol.48, No.2, pp.425-457.［査読あり］

[18] *Masaki Uto (2021) A review of deep-neural automated essay scoring models. Behaviormetrika, Vol.48, No.2, pp.459-484.［査読あり］

[19] *Emiko Tsutsumi, Ryo Kinoshita, Maomi Ueno (2021) Deep Item Response Theory as a Novel Test Theory Based on Deep Learning. Electronics, Vol.10, No.9.［査読あり］

[20] 堤瑛美子，木下涼，＊植野真臣(2021)独立な学習者・項目ネットワークをもつ Deep-IRT．電子情報通信学会論文誌，Vol.J104-D，No.7，pp.596-608.［査読あり］

[21] 内田優斗，＊宇都雅輝（2021）受験者の能力を考慮した深層学習ベース短答記述式問題自動採点手法．教育システム情報学会論文誌 Vol.38，No.3，pp.218-228.［査読あり］

[22] ＊Masaki Uto (2021) Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, Springer, Vol. 53, No. 4, pp. 1440-1454.

[23] ＊Masaki Uto & Maomi Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika, Springer, Vol.47, Issue.2, pp. 469-496.［査読あり］

[24] ○＊Masaki Uto, Duc-Thien Nguyen & Maomi Ueno (2020) Group optimization to maximize peer assessment accuracy using item response theory and integer programming, IEEE

Transactions on Learning Technologies, IEEE Computer Society, Vol.13, No.1, pp.91-106.［査読あり］
[25] ＊宇都雅輝, 植野真臣 (2020) ルーブリック評価における項目反応理論. 電子情報通信学会論文誌 D. Vol.J103, No.05. pp. 459-470.［査読あり］
[26] ＊植野真臣, 木下 涼 (2020) ポスト項目反応理論：深層学習によるテスト理論. Precision Medicine, Vol.3, No.5, pp.56-62.［査読あり］
[27] 木下涼, ＊植野真臣 (2020) 深層学習によるテスト理論：Item Deep Response Theory. 電子情報通信学会論文誌 D, Vol.J103, No.04, pp.314-329.［査読あり］
[28] ＊Yoshimitsu Miyazawa, Maomi Ueno (2020) Computerized Adaptive Testing Method Using Integer Programming to Minimize Item Exposure. Advances in Artificial Intelligence, Vol.1128, pp.105-113.［査読あり］
[29] 渕本壱真, ＊植野真臣 (2020) 等質テスト構成における整数計画法を用いた最大クリーク探索の並列化. 電子情報通信学会論文誌 D, Vol.J103, No.12, pp.881-893.［査読あり］
[30] 八木嵩大・＊宇都雅輝 (2019) パフォーマンス評価における多次元項目反応モデル. 電子情報通信学会論文誌 D. Vol.J102, No.10, pp.708-720.［査読あり］

〔学会発表〕(計 149 件)
[1] *Emiko Tsutsumi, Tetsurou Nishio, Maomi Ueno (2024) Deep-IRT with temporal convolutional network for comprehensive reflection of student ability history data. Artificial Intelligence in Education (AIED).［査読あり］
[2] *Wakaba Kishida, Kazuma Fuchimoto, Yoshimitsu Miyazawa, Maomi Ueno (2023) Item difficulty constrained uniform adaptive testing International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[3] *Misato Yamaura, Itsuki Fukuda, Masaki Uto (2023) Neural automated essay scoring considering logical structure. International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[4] *Takumi Shibata, Masaki Uto (2022) Analytic Automated Essay Scoring based on Deep Neural Networks Integrating Multidimensional Item Response Theory. International Conference on Computational Linguistics (COLING).［査読あり］
[5] *Maomi Ueno, Yoshimitsu Miyazawa (2022) Two-Stage Uniform Adaptive Testing to Balance Measurement Accuracy and Item Exposure, Artificial Intelligence in Education (AIED).［査読あり］
[6] Emiko Tsutsumi, Yiming Guo, Maomi Ueno (2022) Deep knowledge tracing in corporating a hypernetwork with independent student and item networks. International Conference on Educational Data Mining (EDM).［査読あり］
[7] *Masaki Uto (2021) A Multidimensional Item Response Theory Model for Rubric-based Writing Assessment. International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[8] *Itsuki Aomi, Emiko Tsutsumi, Masaki Uto, Maomi Ueno (2021) Integration of Automated Essay Scoring Models using Item Response Theory. International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[9] ＊Masaki Uto, Yikuan Xie & Maomi Ueno (2020) Neural Automated Essay Scoring Incorporating Handcrafted Features. Proceedings of the 28th International Conference on Computational Linguistics (COLING).［査読あり］
[10] ＊Masaki Uto & Masashi Okano (2020) Robust neural automated essay scoring using item response theory. International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[11] ＊Masaki Uto & Yuto Uchida (2020) Automated short-answer grading using deep neural networks and item response theory. International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[12] ＊Maomi Ueno & Yoshimitsu Miyazawa (2019) Uniform adaptive testing using maximum clique algorithm. International Conference on Artificial Intelligence in Education (AIED).［査読あり］
[13] ＊Masaki Uto (2019) Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. International Conference on Artificial Intelligence in Education (AIED).［査読あり］

〔図書〕(計 7 件)
[1] 繁桝算男 (編), 繁桝算男, 加藤健太郎, 光永悠彦, 植野真臣, 宇都雅輝, 二村英幸, 黒田美保 (著)「心理・教育・人事のためのテスト学入門」誠信書房, 2023.