科研費基盤研究(S)
シンポジウム
報告論文集

基盤研究(S) 19H05663「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: 植野真臣
開催日時：2023 年 12 月 27 日（水）
開催方法：ZOOM(オンライン開催)

科研費基盤研究(S) 最終報告書

目次

e テスティングのための自動並行テスト構成手法

渕本 壱真

電気通信大学

1. はじめに

e テスティングとは，異なる問題で構成されるが，同一精度の測定を実現できるコンピュータテストのことである．e テスティングを用いることで，同一能力の受検者が異なるテストを受検しても同一得点となる保証がある．そのために，受検者が同一精度で複数回の受検が可能となるなど様々な利点を持つ．我が国においても医療系共用試験や情報処理技術者試験等が e テスティング上で行われている．また，大学入学試験や公務員試験での導入も検討されており，今後益々e テスティングの需要が高まることが見込まれる．e テスティングでは一般的に，各テストに含まれる出題 項目は異なるが，等質なテスト群が生成される．このようなテストの一つの概念として，項目反応理論を用いた並行テストが知られており，並行テストの自動構成手法が数多く提案されている．

自動並行テスト構成手法の最も重要な課題の一つは可能な限り多くのテストを生成することである．これにより，任意のタイミングでの受検や複数回受検が可能となる．また，これらの自動並行テスト構成手法では各問題項目の出題に偏りが生じる．例えば，この偏りが大きい項目は受検対策により信頼性が低下する．

本資料では，自動並行テスト構成手法のこれらの課題について，本科研費研究で行なってきた研究の概要を紹介する．個別の技術の詳細は，以降に付した原著論文等を参照されたい．

2. テスト数最大化のための自動並行テスト構成

自動並行テスト構成手法の最も重要な課題の一つは可能な限り多くのテストを生成することである．これにより，任意のタイミングでの受検や複数回受検が可能となる．本科研費研究の開始時点では最大クリーク探索を用いた手法が最も多くのテストを生成可能であった．しかし，最大クリーク探索の空間計算量が大きく，最大で 10 万程度のテスト構成が限界であった．そのため，大規模な e テスティングで実用化するためにはテスト数の改善が必要であった．この問題を解決するために，本科研費研究では整数計画法と最大クリーク探索を用いた二段階並列探索手法や zero-suppressed binary decision diagrams（ZDD）を用いた手法の研究を行なってきた[1,3,4,6]．ここでは，これらの研究概要を順に紹介する．

2.1 整数計画法と最大クリーク法を用いた二段階並列探索手法

本研究では最大クリーク探索の空間計算量を緩和するために二段階並列探索手法を提案した[1,4]．第一段階では定数時間で探索可能だが，空間計算量が大きい最大クリーク法でメモリの許す限り多くのテストを生成する．第二段階では空間計算量が小さいが，時間計算量の大きい整数計画法による探索に切り替え，さらに多くのテストを生成する．ただし，整数計画法の時間計算量が大きく，この効果は限定的であるため，第二段階の並列アルゴリズムを考える．具体的には探索中のクリークの全頂点と隣接する頂点の並列探索を一定回数繰り返す．その後，これらの頂点集合から最大クリークを探索し，探索中のクリークと接続する．この提案手法は計算コストを分割することで最大クリーク法よりも多い 1 ヶ月で約 45 万のテストを生成できた．

2.2 Zero-suppressed Binary Decision Diagrams を用いた自動並行テスト構成手法

2-1 の手法は従来手法よりも多くのテストを生成できるが，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（図 1 の右）とは二分決定木（図 2 の左図）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．

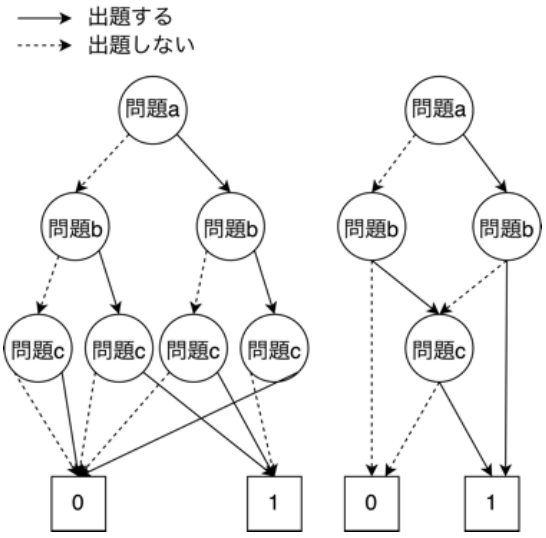

図 1　場合分け二分決定木と ZDD

3. 項目露出ペナルティを用いた整数計画法による自動並行テスト構成

2.の自動並行テスト構成手法では，テスト間に問題項目の重複を許すため，出題頻度（露出数）に偏りが生じさせ，テストの信頼性が低下する．図 2 はこの露出数の偏りを示したものである．図 2 で 4000 回以上出題されているような露出数の大きい問題項目は受験対策され，その項目の信頼性が低下する．また，作問にかかるコストは多大であるため，露出数の小さい問題項目を可能な限り出題することが望ましい．そのため，露出数は可能な限り一様であることが望ましい．

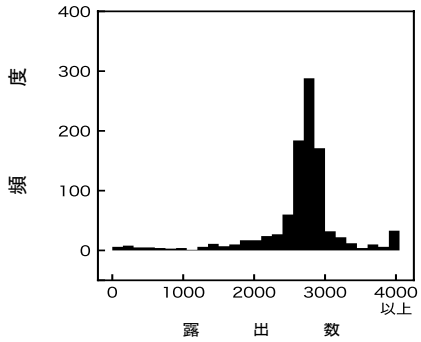

図 2. 露出数の度数分布

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 2-1 の二段階並列探索手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する[7]．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．

業績リスト

1. Kazuma Fuchimoto, Takatoshi Ishii, and Maomi Ueno. Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly. IEEE Transactions on Learning Technologies, 2022. DOI: 10.1109/TLT.2022.3163360.

2. Maomi Ueno, Kazuma Fuchimoto, and Emiko Tsutsumi. E-testing from artificial intelligence approach. Behaviormetrika 48.2, 409-424, 2022. DOI: 10.1007/s41237-021-00143-x.

3. K. Fuchimoto, S. -I. Minato and M. Ueno, "Automated Parallel Test Forms Assembly using Zero-suppressed Binary Decision Diagrams," in IEEE Access, vol. 11, pp. 112804-112813, 2023. DOI: 10.1109/ACCESS.2023.3322720.

4. 渕本 壱真, 植野 真臣. 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-D, No.12, pp. 881-893, 2020. DOI: 10.14923/transinfj.2020JDP7004.

5. 植野 晶, 渕本 壱真, 植野 真臣. 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 2022. DOI: 10.14923/transinfj.2021JDP7037.

6. 渕本 壱真, 湊 真一, 植野 真臣. 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 2022. DOI: 10.1527/tjsai.37-5_A-M23.

7. 渕本 壱真, 植野 真臣. 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 2024. (In Press)

測定誤差と暴露数のトレードオフを調整する適応型テスト

宮澤　芳光[1]

[1]独立行政法人大学入試センター

1. はじめに

近年、テストの結果が受検者に大きな影響を与えるハイ・ステークスなテストでeテスティングが実用化されつつある。eテスティングとは、Web上でテストを受けるCBT（Computer based testing）であり、受検者が何度でも等質な測定誤差で異なる項目から構成されたテストを受検できる。e テスティングには、適応型テスト（CAT: Computerized Adaptive Testing）と呼ばれるテスト出題方式が知られている。適応型テストは、受検者の能力値を逐次的に推定し、その能力値に応じて情報量が最大の項目を出題することで、測定誤差を増加させずにテストの長さや受験時間を短縮できる。しかし、同一の受検者が複数回受験した場合には、同一の項目群が出題される傾向があり、実際に適応型テストを導入しているテスト業者の重要な問題になっている。また、特定の項目群が多くの受検者に対して提示されてしまうため、別の受検者に共有されてしまうことが指摘されている。これらの問題を解決するため、露出制御を用いた適応型テストが多数提案されている。代表的な手法として、van der Linden(2010)らは、露出数の制約を用いて項目集合（シャドーテストと呼ばれる）を逐次的に構成し、その項目集合から項目選択する手法を提案している。この手法では、項目選択のたびにシャドーテストを構成し、シャドーテスト中の項目から情報量が最大の項目を選択する。一方、別のアプローチとして、確率的に項目選択を制御する手法が提案されている。Sympson-Hetter 手法(Hetter & Sympson 1997)では、事前にシミュレーション実験を用いて項目の出題確率とその出題確率を制御するパラメータを算出し、そのパラメータを用いて確率的に項目の出題を制御する手法が提案されている。これらの手法では、特定の項目群の過度な露出を防ぐことはできる。しかし、これらの手法では、露出数の一様性は担保されるが、情報量を制約しているために測定誤差が増加してしまう問題点がある。すなわち、露出数の減少と測定誤差の増加には、トレードオフの関係がある。

本稿では、この問題を解決するために、露出数の減少と測定誤差の増加のトレードオフを制御する等質適応型テストを紹介する。

2. 等質適応型テスト

適応型テストでは、露出数の減少と測定誤差の増加にはトレードオフの関係がある。本研究では、暴露数の減少と測定誤差の増加のトレードオフを制御するために，等質適応型テスト（Uniform Adaptive Testing; 以降 UAT と呼ぶ）を提案してきた。UAT は，最新の等質テスト構成技術を用いて情報量が等質な項目集合にアイテムバンクを分割し，受検者ごとに異なる等質テストを割り当ててアイテムバンクとみなして適応型テストを実施する。受検者ごとに異なる項目集合をアイテムバンクとして用いるため，能力値が同等な受検者であっても異なる項目群を出題することができ，受検者間の測定誤差を等質にしつつ，暴露数を減少できる。等質テストとは、複数のテストが異なる項目で構成されているにも関わらず、等質な情報量を持つ項目集合である。等質テスト構成技術は、既にeテスティング技術の基幹技術として多数研究されている。当初は、その計算量の大きさから少数の等質テストしか生成できなかった。しかし，近年の研究では，大規模な数の等質テストを生成できる技術が開発され、医療系共用試験などの実際のテスト運営でも実用化されている。等質適応型テストでは、当時、最大数の等質テストを構成できるIshiiらの手法(Ishii et al. 2014)を用いてアイテムバンクを多数の項目集合に分割し、受検者ごとに異なる項目集合を割り当て、その項目集合から情報量が最大の項目を選択している。これにより，暴露数と測定誤差のトレードオフを制御でき、さらに，等質適応型

テストが過学習を避け，テストの長さを短縮できた。ここでは、等質適応型テストの項目選択アルゴリズムを紹介する。

等質適応型テストでは、まず、複数等質テスト構成手法を用いて，情報量や回答所要時間等の統計的性質は等質であるが，異なる項目から構成された等質テストを複数生成する。複数等質テストを生成できる手法としては，Ishii らの手法(Ishii et al. 2014)が知られている．Ishii らの手法では，テスト構成問題を最大クリーク問題として扱う．具体的には，次のグラフを考え，そこから最大のクリーク（その集合に含まれる任意の頂点がすべて結合されている）構造を抽出することで複数等質テストを構成する．

頂点:与えられたアイテムバンクから構成可能な，重複条件以外の全てのテスト構成条件を満たす，可能テストを頂点とする．

エッジ:二つの可能テストが重複条件を満たしている場合（重複条件により指示される最大重複項目数より少ない重複項目しか持っていないなら）その二つの頂点（テスト）間にエッジを張る．

このように作成されたグラフのクリークは所望の等質条件を満たした等質テストの集合と解釈できる．そのため，このグラフの最大クリークを抽出することで，最大数の複数等質テストを生成できる．本研究では，受検者ごとにできる限り異なる等質テストを割り当てるため，非常に多くの等質等質テストを構成する必要がある．そこで，本研究では，等質テスト構成手法として Ishii らの手法を用いる．

UAT では、Ishii らの手法により生成された複数等質テストの中から，各受検者に異なるテストを一つ割り当て，それに含まれる項目集合をアイテムバンクとみなして適応型テストを行う．具体的には、以下の通りである。

1. 受検者の能力値を初期化する．
2. 能力値を所与として，フィッシャー情報量が最大となる項目を割り当てられた等質テストから選択し，出題する．
3. 項目に対する正誤データから受検者の能力推定値を更新する．
4. 上記の手順 2 と 3 を，受検者の能力推定値の更新幅が閾値以下になるまで繰り返す．

UAT では，受検者ごとに異なる項目集合から項目を選択するため，能力が同等な受検者であっても異なる項目が出題できる．さらに，出題される項目の多様性が向上するため，アイテムバンク内の項目を満遍なく利用することができ，露出数の偏りも軽減される．

以上の研究の成果は、国際会議 AIED に論文が採択されている[5]。

3. 2 段階等質適応型テスト

UAT では、暴露数と測定誤差のトレードオフを制御でき、さらに、テストの長さを短縮できたことを報告している．しかし，一般的にテストの長さが短縮すると能力推定値の測定誤差が大きくなることが多い．UAT の評価実験では、従来の適応型テストの収束基準に従い，受検者の能力推定値が収束することでテストを終了させ，フィッシャー情報量の平方根の逆数により求められる漸近誤差は評価しているが，受検者の真の能力値と能力推定値の差異による測定誤差に関して評価は行っていない．分析の結果、アイテムバンクを分割することで項目候補が少なくなり，枯渇して見かけ上能力推定値が収束していた．このため，能力真値と能力推定値が大きく乖離している．ここでは、UAT の問題を解決するため，2 段階等質適応型テストを紹介する。

2 段階等質適応型テストでは，事前にアイテムバンクを分割して情報量が等質な項目集合を複数構成し，テストの前半に項目集合から項目選択し，受検者の能力推定値が収束し始めたテストの後半にアイテムバンク全体から項目選択する．アイテムバンクの分割には，当時、世界最大数の等質テスト構成を可能にする石井らの手法[4]を用いる．石井らでは，整数計画問題を用いて最大クリークを逐次探索し，効率的に等質テストを構成できる手法を提案している．

石井他(2017)の手法[4]では，構成中の等質テスト群を C，構成済みの等質テスト数を|C|とし，以下の整数計画問題を用いて等質テストを構成する．

$$Maximize \sum_{i=1}^{N} \lambda_i x_i$$

Subject to

$$\sum_{i=1}^{N} y_{ki} x_i \leq O\left(\text{項目の重複上限数}\right); (k = 1, \cdots, |C|)$$

$$\sum_{i=1}^{N} x_i = n\left(\text{テストの長さ}\right)$$

$$\sum_{i=1}^{N} I_i(\theta_l) x_i = I(\theta_l)$$

$$LB(\theta_l) \leq I(\theta_l) \leq UB(\theta_l)$$
$$(l = 1, \cdots, L)$$

$$x_i = \begin{cases} 1\text{: 項目}i\text{が等質テストに含まれるとき,} \\ 0\text{: 上記以外} \end{cases}$$

$$y_{ki} = \begin{cases} 1\text{: }i\text{番目の項目が}C\text{中の}k\text{番目の等質テストに含まれるとき,} \\ 0\text{: 上記以外} \end{cases}$$

ここで，λ_iは，互いに独立な[0, 1)の連続一様分布からの乱数であり，本問題が解かれるたびにリサンプリングされる．構成済みの等質テストをCとして，Oは，重複項目数の上限である$LB(\theta_l)$は，情報量の下限であり，$UB(\theta_l)$が上限である．$LB(\theta_l)$と$UB(\theta_l)$は，アイテムバンクに含まれる項目の特性に応じて適切に設定する必要がある．本研究では，項目の情報量の平均をm_{ib}とし，標準偏差をsd_{ib}，等質テストのテストの長さをnとしたとき，情報量の上限を$(m_{ib} + sd_{ib})n$とし，下限を$m_{ib}n$とした．

2 段階等質適応型テストでは，事前に石井他(2017)の手法を用いてアイテムバンクを分割して情報量が等質な項目集合を複数構成する．この項目集合を用いた 2 段階等質適応型テストのアルゴリズムについて詳述する．第 1 段階では以下のアルゴリズムに従って項目選択する．

1. 項目集合をランダムに割り当てる．
2. 能力推定値を$\hat{\theta} = 0$に初期化する．
3. 能力推定値$\hat{\theta}$を所与として項目集合から情報量が最大となる項目を選択して出題する．
4. 項目への反応データとそれまでの解答履歴から能力推定値$\hat{\theta}$を求める．
5. 手順(3)と(4)を$\hat{\theta}$の更新幅が閾値以下になるまで繰り返す．

次に，第 2 段階では以下のように出題方略が変更される．

1. 能力推定値$\hat{\theta}$を所与としてアイテムバンク全ての項目集合から情報量が最も高い項目を選択する．ハイステークスな試験では，暴露数の上限値が制約として決まっている場合がある．必要に応じて，このステップで暴露数の上限値を制約として組み込む．
2. 項目への反応データとそれまでの解答履歴から能力推定値$\hat{\theta}$を求める．
3. 手順(1)と(2)をテストの終了条件まで繰り返す．

従来の適応型テストでは，能力推定値の更新幅が閾値以下になるときをテスト終了条件として設定していた．しかし，等質適応型テストでは，テスト終了条件を能力推定値の更新幅が閾値以下とした

場合，項目集合に情報量の高い項目がなくなり，能力推定値が能力真値に収束する前にテストが終了することがある．そこで本研究では，テストの長さをテスト終了条件とした．本手法では，上記の手順でアイテムバンクの項目をできる限り一様に活用しながら受検者の能力を高精度で推定できる．

以上の研究の成果は、日本テスト学会で大会発表賞を受賞し[6]、国際会議 AIED に論文が採択され[3]、電子情報通信学会の論文誌に採択されている[1]。

4. 項目難易度制約付き 2 段階等質適応型テスト

2 段階等質適応型テストでは、従来の適応型テストと同等の測定精度を保ちつつ，従来手法よりも暴露数を減少できた.しかし，TUAT は識別力パラメータの大きい項目に偏って暴露されるという問題がある.UAT は、1 段階目で推定値が能力真値の近傍に収束するため，2 段階目ではその能力推定値近傍の難易度をもつ項目のみが出題され，暴露数の偏りは軽減できると仮定している.この仮定は，一つの項目については，その難易度に一致する能力値のフィッシャー情報量が最大となることを根拠としている.しかし, 実際は項目の難易度が能力推定値と乖離していても, 乖離していない場合に比較して, 情報量が大きくなる場合が多くある.ここでは、TUAT の問題を解決するため，項目難易度制約付き 2 段階等質適応型テストを紹介する。

項目難易度制約付き 2 段階等質適応型テストでは、TUAT の 2 段階目以降は能力推定値近傍の難易度パラメータをもつ項目に限定して出題する。2 段階目では，能力推定値近傍の難易度パラメータ区間に属する項目から適応的項目出題を行う.その難易度パラメータ区間は, 次式に示す能力推定値の事後標準偏差に比例して決定する.

$$SD(\hat{\theta}) = \sqrt{\int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta|\boldsymbol{u})}$$

本手法では，次の区間を能力推定値近傍とした.

$$\hat{\theta} - \delta SD(\hat{\theta}) < \theta < \hat{\theta} + \delta SD(\hat{\theta})$$

ただし，δ は難易度パラメータ区間に対する事後標準偏差の影響度合いを決定する重み付けチューニングパラメータである.能力推定値近傍の項目の難易度パラメータ区間は次の通りである

$$\hat{\theta} - \delta SD(\hat{\theta}) < b < \hat{\theta} + \delta SD(\hat{\theta})$$

項目難易度制約付き 2 段階等質適応型テストでは、次のアルゴリズムにしたがって項目出題する.

(1) 各受検者に等質アイテムバンクをランダムに 1 つ割り当てる
(2) 能力推定値を$\hat{\theta} = 0$に初期化する.
(3) 割り当てられた等質アイテムバンクから情報量が最大の項目を出題する.
(4) 反応データから能力推定値$\hat{\theta}$を更新する.
(5) 能力推定値$\hat{\theta}$の更新幅が閾値ε未満になるまで手順(3)，(4)を繰り返す.

能力推定値$\hat{\theta}$の更新幅が閾値ε未満になった後，第 2 段階に移行する.

(1) 能力推定値$\hat{\theta}$と事後標準偏差 SD($\hat{\theta}$)から難易度パラメータ区間を計算する.
(2) 暴露数の上限が設定されている場合は, 暴露数が上限に達した項目をアイテムバンクから除く.
(3) 難易度パラメータ区間に属する項目から情報量が最大の項目を出題する.
(4) 反応データから能力推定値$\hat{\theta}$を更新する.
(5) 能力推定値$\hat{\theta}$の更新幅が閾値ε未満になるまで手順(1)から(4)を繰り返す.

本手法により，2 段階目以降は能力推定値近傍の難易度パラメータをもつ項目に限定して出題でき，従来手法と比較して測定精度を同等に保ちつつ，暴露数の偏りを減少させることができる。

以上の研究の成果は、国際会議 AIED に論文が採択されている[2]。

5. むすび

本原稿では、基盤研究 S「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」で得られた成果の概要を報告した。各章で紹介した論文の詳細については以下の業績リストの通りである。

引用文献

van der Linden, W. J., & Glas, C. A. (2010). Elements of adaptive testing, Springer.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 141–144). American Psychological Association.

T. Ishii, P. Songmuang, and M. 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.

石井隆稔･赤倉貴子･植野真臣 (2018). 複数等質テスト構成に おける整数計画問題を用いた最大クリーク探索の近似法, 電子情報通信学会論文誌.D, vol.J101, pp.725‐728.

業績リスト

【査読付き論文誌】

[1] 宮澤芳光・植野真臣 (2023) 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-D, No.1, pp. 34‐46.

【国際会議発表】

[2] Wakaba Kishida, Kazuma Fuchimoto, Yoshimitsu Miyazawa, Maomi Ueno (2023) Item difficulty constrained uniform adaptive testing. International Conference on Artificial Intelligence in Education (AIED), Communications in Computer and Information Science, vol. 1831, pp. 568–573.

[3] Maomi Ueno, Yoshimitsu Miyazawa (2022) Two-Stage Uniform Adaptive Testing to Balance Measurement Accuracy and Item Exposure. International Conference on Artificial Intelligence in Education (AIED), Lecture Notes in Computer Science, vol. 13355, pp. 626–632.

[4] Yoshimitsu Miyazawa, Maomi Ueno (2020) Computerized Adaptive Testing Method Using Integer Programming to Minimize Item Exposure. Advances in Artificial Intelligence: Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI), pp.105–113.

[5] Maomi Ueno, Yoshimitsu Miyazawa (2019) Uniform adaptive testing using maximum clique algorithm. International Conference on Artificial Intelligence in Education (AIED), Lecture Notes in Computer Science, vol. 11625, pp. 482–493.

【受賞】

[6] 日本テスト学会大会発表賞 (2022）宮澤 芳光，渕本　壱真，植野 真臣．「等質テスト構成の並列化技術を用いた 2 段階等質適応型テスト」

[7] JSAI Excellence Award: International Session (2019) Yoshimitsu Miyazawa, Maomi Ueno, Computerized Adaptive Testing Method using Integer Programming to Minimize Item Exposure.

適応的学習支援のための学習者の反応予測とパラメータの解釈性を両立する Deep-IRT の開発

堤　瑛美子

東京大学　情報理工学系研究科

1. はじめに

近年，教育現場ではコンピュータやタブレット端末の普及に伴ってオンライン学習システムを用いた学習が広まり，大量の教育ビッグデータ（学習者が課題に取り組んだ日時や回答）を如何に有効活用するかが課題になっている．特に，学習支援システム分野や人工知能分野では機械学習を用いて教育ビッグデータを分析することにより学習者の特性や成長に合わせて適切な問題提供，学習支援を行うアダプティブラーニングが注目されている．具体的には，学習履歴データから学習過程における学習者の習熟度（理解度）の変化と未知の課題への反応(正答・誤答)を予測することで得意分野・苦手分野を把握し，個人に適切な学習支援を行う．教育における最も難しい問題とは，教師が学習者に教えすぎると，学習者の十分な発達は望めないということである．そのため，学習者の自立を促すためには，学習者が自力で解決できる適度な支援を行う必要があると考えられている．したがって，学習者の成長を正確に把握し，未知の課題への反応を正確に予測することがアダプティブラーニングにおける重要な課題となる．

これまで，既存の確率モデルを用いたアダプティブラーニングシステムの開発が盛んに行われてきた．代表的な確率モデルの一つに項目反応理論（Item Response Theory: IRT）[1]が挙げられる．IRT は学習者の過去の学習データをもとに現在の習熟度を推定する数理モデルであり，学習者の習熟度や課題の特性を表す解釈性の高いパラメータをもつため，教育的解釈性の高いモデルとしてさまざまな現場で利用されている．しかし，従来の IRT はテスト理論の中で開発されており，学習者の習熟度が固定されているために学習過程を反映できず，学習支援システムへの応用は限定的であった．また，オンライン学習システムで収集される教育ビッグデータはスパースなデータが多く，学習者のランダムサンプリングが保証されていない．そのため，従来の確率モデルではパラメータの解釈可能性は非常に高い反面，学習者の高精度な習熟度推定は難しいことがわかってきた．この問題を解決するために，近年では深層学習を教育ビッグデータに適応し，学習過程で時系列変化する学習者の習熟度を推定しながら課題への反応予測を行う手法が提案されている．しかし，一般に深層学習手法は高い反応予測精度を示す反面，教育的な意味でのパラメータの解釈性をもたず教育応用には限界がある．したがって，確率モデルにおけるパラメータの解釈性と深層学習手法を用いた高精度な反応予測の両立が課題となっている．

本資料では，適応的学習支援のための学習者の反応予測とパラメータの解釈性を両立する手法について，本科研費研究で発表者らが行った研究の概要を紹介する．

2. テスト理論のための Deep-IRT

申請者らは初めに，学習過程で学習者の能力値が一定であることを仮定し，深層学習を用いたテスト理論のための Deep-IRT を提案した．一般にテスト理論では，あるテストの結果に対する学習者の

能力評価を行う手法として IRT が用いられている．IRT の特徴は，学習者の同一母集団からの独立ランダムサンプリングを仮定し，異なるテストを受検した学習者を同一尺度上で評価できる点である．しかし，実際の教育現場では，学習者の能力値が標準正規分布からランダムサンプリングされる保証はなく，偏ったテストデータがサンプルされることが多いという問題があった．一方，提案手法は学習者の独立ランダムサンプリングを仮定せず，学習者間や項目間の関係性を考慮しながらパラメータ推定を行うことができる．

本研究では，学習者の学習履歴データから二つの独立なニューラルネットワークを用いて能力パラメータと項目難易度パラメータを推定することで，IRT と同等のパラメータ解釈性を実現する Deep-IRT を開発した．一般的に，複数の独立なニューラルネットワークを用いた手法は汎化性能が低下する問題があるが，本手法はネットワークの深層化およびドロップアウトの技術を用いることで IRT を上回る反応予測と能力値推定を達成した． 本手法は学習者の能力値が学習過程で固定されている状況を仮定したものでテスト理論のための手法である．

【本研究の詳細は 1 を参照されたい．】

3 適応的学習支援のための Deep-IRT

3.1 多次元の能力時系列変化を推定する Deep-IRT

適応的学習支援では学習過程における学習者の能力成長を時系列に追跡する必要がある．さらに，学習範囲ごとに獲得するべきスキル(技能)が複数存在しており，学習者の成長予測には全てのスキルに対する能力変化を考慮しなければならない．

人工知能分野では，パラメータの解釈性と高精度な反応予測を両立するために，深層学習手法と IRT の枠組みを組み合わせた Deep-IRT 手法が複数開発されてきた [2,3]．しかし，既存手法は学習者の能力の時系列変化を考慮していない手法や，能力値パラメータが課題の難易度パラメータに依存して推定されるために解釈性が低下している手法であり，依然として教育応用には課題があった．

そこで申請者らは，学習者の能力パラメータと課題・スキルの難易度パラメータをそれぞれ独立したネットワークを用いて推定し，多次元スキルに対する能力値の時系列変化を推定する新たな Deep-IRT を提案した．提案手法では多次元の能力値を保存する潜在変数をもち，毎時点に潜在変数の更新を行う．具体的には，それまでの潜在変数の値をどの程度保存し，最新の課題への反応データをどの程度反映するか調整する忘却パラメータを推定することで潜在変数の更新を行う．これらの更新・忘却機能により，提案手法は深層学習手法の反応予測精度を保ちつつ，能力パラメータの解釈性を向上させた．

【本研究の詳細は業績リスト 2,6,7 を参照されたい．】

3.2 能力の忘却を最適化する Hypernetwork を組み込んだ Deep-IRT

提案手法（業績リスト 2,6,7）では，学習者の能力パラメータと課題の難易度パラメータの解釈性を向上させたが，学習者の反応予測精度については既存手法と同程度であった．これは，モデルが多次元能力を保持する潜在変数を更新する際に最新の反応データのみを用いるために，学習者の過去の反応データや潜在能力を考慮できず，反応予測精度の向上を妨げている可能性があった．

申請者らは，この問題を解決するために，本研究では Tsutsumi らの Deep-IRT （業績リスト 2,6,7）に新たな Hypernetwork を組み込み，最新の学習者の反応データと直前の学習者の潜在能力値の両方を用いて忘却パラメータを最適化した．Hypernetwork は，近年，自然言語処理の分野で LSTM を拡張するために用いられている方法である[4]．一般的な LSTM では潜在変数を更新するための重みパラメータの値が全時点で共有されているが，Hypernetwork を用いて潜在変数を更新する前に重みを時点

ごとに最適化することで，LSTM のパフォーマンスが向上することが示されている[4]．本研究では，この手法を応用し，Deep-IRT における潜在変数を更新する前に，Hypernetwork 内で最新の学習者の反応データと前時点での潜在変数のバランスを調整しながら忘却パラメータを推定する．これにより，過去の反応データの適切に反映した能力推定と反応予測が可能となる．

Hypernetwork を用いた Deep-IRT 手法は，最先端の反応予測手法を上回る予測精度を示した．特に，忘却パラメータ最適化は長期の学習において有効であることがわかった．さらに，多次元のスキルに対する学習者の能力推移を可視化し，先行研究における能力推定を精度と解釈性を改善することを示した．

【本研究の詳細は業績リスト 3,4,7 を参考にされたい．また，本論文は電子情報通信学会において 2023 年度論文賞を受賞した．】

3.3　TCN を組み込んだ学習者の長期の能力変化を反映する Deep-IRT

前章で紹介した Deep-IRT（業績リスト 3,4,7）は，既存手法の反応予測精度と能力パラメータの推定精度を向上させることを示した．一方で，未知の課題への正答確率を計算する際に使用する能力パラメータを現時点での潜在変数のみから推定している．そのため，過去の能力変化を十分に考慮した能力パラメータの推定が行われていない可能性があった．

そこで申請者らは，毎時点に多次元の潜在変数から出力された多次元の能力状態を保存し，Temporal Convolutional Network (TCN) [5]で過去の能力状態を畳み込むことによって，学習者の長期の能力変化を反応予測に反映する新たな Deep-IRT 手法を提案した．TCN は時系列データを予測する分野で使われているニューラルネットワークモデルであり，近年，LSTM や GRU などの RNN ベースのモデルよりも高い精度で特定の時系列データを予測することが知られている[5]．複数の時系列データを畳み込む TCN では，より長期記憶性をもつことが示されている．さらに，TCN は長期間における時系列データや時系列変化する潜在変数を多層で畳み込むことで，より広範囲の時系列データの特徴を出力に反映させることができる．また，Causal Dilated Convolution と Residual Connection と呼ばれる手法を用いることで，より長期的な学習データに対してパラメータ数の増加を抑えながら反応予測を行うことが可能である．

提案手法は既存の Tsutsumi らの Deep-IRT 手法（業績リスト 3,4,7）の反応予測精度を上回る精度を達成した．また，シミュレーションデータを用いた能力推定精度比較では，真の能力が一時点前の能力に依存するような学習過程の場合において，提案手法が既存手法より高精度に能力推定を行うことを示し，TCN が有効に機能していることを示した．

【本研究の詳細は 5,8 を参考にされたい．また，本論文(業績リスト 5)は電子情報通信学会において 2023 年度論文賞を受賞した．】

1　むすび

本研究では，主に学習者の反応予測を高精度に行い，パラメータの高い解釈性をもつ Deep-IRT 手法の開発を行った．提案手法はこれまでの既存手法を上回る精度を示し，適応的学習支援への応用が期待できる．研究成果は IEEE，電子情報通信学会などのトップジャーナルや教育工学分野でのトップカンファレンス AIED に採録されている．

引用文献

[1] M. Ueno and Y. Miyazawa, :IRT-based adaptive hints to scaffold learning in programming, IEEE Transactions on Learning Technologies, IEEE computer Society, vol.11, Issue 4, pp.415-428 ,2018.

[2] G. Converse, S. Pu, S. Oliveira,: Incorporating Item Response Theory into Knowledge Tracing, Artificial Intelligence in Education. AIED 2021.

[3] C. Yeung,: Deep-irt: Make deep learning based knowledge tracing explainable using item response theory, Proceedings of the 12th International Conference on Educational Data Mining, EDM,2019.

[4] D. Ha, A. Dai, and Q.V. Le, : Hypernetworks, arXiv preprintarXiv:1609.09106, 2016.

[5] S. Bai, J.Z. Kolter, and V. Koltun,: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018. https://arxiv.org/abs/1803.01271

業績リスト

【査読付き論文誌】

1. Emiko Tsutsumi, Ryo Kinoshita and Maomi Ueno,:Deep Item Response Theory as a Novel Test Theory Based on Deep Learning.Electronics,no.10, pp.1020,2021, DOI:10.3390/electronics10091020.
2. 堤瑛美子，木下涼，植野真臣，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 D, Vol.J104, No.7, pp.596-608, 2021, DOI: 10.14923/transinfj.2020JDP7061
3. 堤 瑛美子, 郭 亦鳴, 植野 真臣, 学習データの忘却を最適化する Hypernetwork を組み込んだ DeepIRT，電子情報通信学会論文誌 D, Vol.J106-D,No.02,pp.-,Feb. 2023, 10.14923/transinfj.2022LEP0003
4. E. Tsutsumi, Y. Guo, R. Kinoshita and M. Ueno, “Deep Knowledge Tracing Incorporating a Hypernetwork With Independent Student and Item Networks,” in *IEEE Transactions on Learning Technologies*, vol. 17, pp. 951-965, 2024, doi: 10.1109/TLT.2023.3346671.
5. 西尾 徹朗 , 堤 瑛美子 , 植野 真臣, 学習者の能力の時系列変化を畳み込む Temporal Convolutional Network を組み込んだ Deep-IRT，電子情報通信学会論文誌 D, Vol.J107-D,No.03,pp.-,Mar. 2024, 10.14923/transinfj.2023JDP7018

【国際会議発表】

6. Emiko Tsutsumi, Ryo Kinoshita and Maomi Ueno,:Deep-IRT with independent student and item networks, International Conference on Educational Data Mining, EDM, 2021.
7. Emiko Tsutsumi, Yiming Guo, Maomi Ueno,: Deep knowledge tracing incorporating a hypernetwork with independent student and item networks, International Conference on Educational Data Mining, EDM, 2022.
8. Emiko Tsutsumi, Tetsurou Nishio, Maomi Ueno, Deep-IRT with temporal convolutional network for comprehensive reflection of student ability history data, 25th International Conference, AIED 2023.

【受賞】

1. 電子情報通信学会 2023 年度論文賞
2. 2021 年度人工知能学会全国大会 大会優秀賞
3. 2021 年度教育システム情報学会 大会奨励賞
4. 2019 年度人工知能学会全国大会 大会優秀賞

パフォーマンス評価のための項目反応理論と記述式回答自動採点，および問題自動生成

宇都雅輝

電気通信大学

1. はじめに

近年，様々な学習・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．パフォーマンス評価は，実践的・現実的な課題に対する受検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，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様々な形式で活用されてきた．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と 2）採点コストの高さよる大規模試験実施の困難さ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検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．これらのモデルは，テスト理論の一つである項目反応理論（Item Response Theory: IRT）に基づくモデルとして定式化されており，様々なパフォーマンス・テストの分析や信頼性改善に利用されてきた．

2）の問題を解決するアプローチとしては，自動採点技術の活用が注目されている．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おり，現在も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いる．深層学習に基づく自動採点技術は，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毎年新たな提案がなされ，精度が更新され続けている．

本資料では，パフォーマンス評価のための項目反応理論と記述・論述式試験の自動採点技術について，基盤研究 S「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」の支援を受けて我々が行ってきた研究の概要を紹介する．

さらに，e テスティングの運用においてコストの高い作業の一つとして「作問」が挙げられる．近年では自然言語処理技術の発展により，自動作問技術の高度化も進展している．我々も，英語の読解課題を主たる対象として問題生成技術の研究を進めている．本資料では，そのような自動作問の研究成果についても紹介する．

2. パフォーマンス評価のための項目反応理論

パフォーマンス評価では，評価結果が評価者や課題の特性に強く依存する問題があり，これが能力測定の信頼性を低下させる要因となることが知られている．評価のバイアス要因となる代表的な評価者特性としては，甘さ・厳しさ（Leniency/Severity）や一貫性（Consistency），尺度範囲の制限（Restriction of Range）などが知られており，課題特性としては困難度（Difficulty）や識別力（Discrimination）の影響が大きいとされてきた．

この問題を解決する手法の一つとして，評価者と課題の特性を考慮して受検者の能力を推定できる項目反応モデルが近年多数提案されている．これらのモデルでは，評価者と課題のバイアスを考慮して受検者の能力を推定できるため，合計点や平均点といった単純な得点化法よりも高精度な能力測定が可能となる．評価者と課題の特性を考慮した代表的な項目反応モデルとしては，Linacre が提案した多相ラッシュモデルが広く知られている．他方で，多様な評価者特性・課題特性の影響が想定される

場合，多相ラッシュモデルではそれらの特性を十分に表現できず，能力測定精度が低下してしまう．この問題を解決するために，我々は評価者や課題の多様な特性を考慮できる多相ラッシュモデルの拡張モデルとして，一般化多相ラッシュモデルを提案した．

2.1 一般化多相ラッシュモデル

一般化多相ラッシュモデルでは，課題 i への受検者 j の回答に評価者 r が評点 k を与える確率 $P(x_{ijr}=k)$を次式で定義する．

$$P(x_{ijr}=k)=\frac{exp\sum_{m=1}^{k}(\alpha_i\alpha_r(\theta_j-\beta_i-\beta_r-d_{rm}))}{\sum_{l=1}^{K}exp\sum_{m=1}^{l}(\alpha_i\alpha_r(\theta_j-\beta_i-\beta_r-d_{rm}))}$$

ここで，α_iは課題 i の識別力，α_rは評価者rの評価の一貫性，θ_jは受検者jの能力，β_iは課題iの困難度，β_rは評価者r の厳しさ，d_{rm}は評価カテゴリ m に対する評価者rの厳しさを表す．なお，モデルの識別性のために以下の制約を課す．

$$\theta_j\sim N(0,\ 1),\ \prod_r\alpha_r=1,\ \sum_r\beta_r=1, d_{r1}=0,\ \sum_m d_{rm}\ =0$$

従来の多相ラッシュモデルでは評価者の厳しさと課題困難度しか考慮できなかったのに対し，このモデルでは，課題の識別力，および評価者の一貫性と尺度範囲の制限（特定の評価カテゴリを過剰に使用する傾向）も表現できるため，多様な評価者特性・課題特性の影響が想定される場合に多相ラッシュモデルよりデータに柔軟に適合し，より高精度な能力測定を実現できる．

実データ実験の結果，提案モデルがモデル適合度・能力推定精度ともに従来モデルよりも高い性能を示したことを確認した．また，提案モデルには，モデル適合度と能力測定精度の改善に加えて，各評価者や課題の特性をより詳細に分析できるという利点も有する．

本研究の成果は，国際論文誌 Behaviormetrika に採択された[14].

2.2 拡張モデル

本科研費研究では，一般化多相ラッシュモデルを基礎モデルとして様々な拡張モデルの開発も行った．本節では，ルーブリックを用いた評価のための拡張モデルと評価者特性の時間変動を考慮した時系列型モデルへの拡張を紹介する．

2.2.1 ルーブリックを用いた評価のための拡張モデル

ルーブリックを用いた評価では，一般に複数の評価観点に基づいた採点が行われるため，評価結果は評価者と課題だけでなく，評価観点の特性にも依存する．そこで，我々は課題と評価者の特性に加えて，評価観点の特性も同時に考慮できるモデルとして，一般化多相ラッシュモデルに評価観点の特性を表すパラメータを追加したモデルを提案した．この拡張モデルでは，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推定が可能であるため，より高精度な能力測定が可能になる．

さらに，ルーブリックで定義される評価観点は複数の次元の能力を測定していると想定できる場合がある．例えば，ライティング能力を測定するルーブリックは，論理性や独創性，表現力などの複数の下位能力を測定する評価観点で構成されることが多い．しかし，一般化多相ラッシュモデルを含む従来の項目反応モデルは，測定対象の能力が単一であることを意味する能力の 1 次元性を仮定しており，多次元での能力測定には利用できなかった．そこで我々は，多次元型の項目反応モデルの一つである多次元段階反応モデルを拡張し，評価者と評価観点の特性を考慮して多次元尺度で能力を推定できるモデルも提案した．このモデルを適用することで，ルーブリックのどの評価項目が共通した下位能力を測定しているかを読み取ることができ，対象のルーブリックが測定している能力の構成概念を分析できる．さらに，そのように得られる多次元尺度上で各受検者の能力を推定することができる．

以上の研究の成果は，電子情報通信学会論文誌と国際論文誌 Behaviormetrika，国際会議 AIED に複数の論文が採択された[5, 10, 16, 18, 29]．また，国際会議 AIED では Best paper にノミネートされ，国内学会では人工知能学会で受賞をしている[受賞欄 26].

2.2.2 評価者特性の時間変動を考慮した時系列モデル

評価者特性を考慮した従来モデルのほとんどは採点過程で評価者の厳しさが変化しないことを仮定している．しかし，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，厳しさが採点の過程で変化する「評価者特性ドリフト」と呼ばれる現象が生じる場合がある．そこで我々はそのような評価者特性ドリフトをとらえることができるモデルを提案した．具体的には各評価者の採点結果データを一定の時間幅を持つ区分に分割し，それぞれの時間区分ごとに評価者の厳しさを推定できるモデルを開発した．さらに，提案モデルでは，各時間区分における評価者の厳しさが時間的に依存することを考慮するために，評価者の厳しさパラメータがマルコフモデルに従うように設計されている．これにより，時間区分ごとの厳しさを独立に推定するよりも，より安定的に厳しさの推定値が得られる．さらに，評価者特性ドリフトの大きさに関する分析者の主観も考慮できるように，提案モデルは階層ベイズモデルとして定式化されている．実データ実験の結果，本手法により評価者特性ドリフトを適切に捉えられることが示された．

本研究の成果は，国際論文誌 Behavior Research Methods に採択された[4]．また，国内学会では教育システム情報学会で受賞をしている[受賞欄 17].

2.3 評価者割り当てへの応用

評価者と課題の特性を考慮した項目反応理論では，各評価者が個別の受検者の能力をどの程度の精度で測定できるかを「情報量」という概念で推定できる．項目反応理論で利用される代表的な情報量であるフィッシャー情報量は，その逆平方根の二乗が能力測定の標準誤差の推定値となるため，能力測定の精度を表す指標として解釈できる．ある受検者に対してフィッシャー情報量が高い評価者ほどその受検者の能力を正確に評価できるとみなせる．そこで，我々は，多数の評価者が分担して採点を行うような大規模試験において，フィッシャー情報量を最大化するように各受検者に最適な評価者を割り当てる手法を開発した．具体的には，各受検者に対する情報量を最大化する整数計画問題として評価者割り当て問題を定式化した．この手法で評価者割り当てを最適化することで，効率的に能力測定の精度を改善できることを示した．

本研究の成果は，国際論文誌 IEEE Transactions on Learning Technologies に採択された[17].

2.4 言語処理技術を活用した能力測定精度の改善

評価者と課題の特性を考慮した項目反応モデルは，平均点などの単純な手法と比べて一般に高精度な能力測定を実現できる．しかし，受検者あたりの評価者数が極端に少ない場合には，このようなモデルを利用しても能力測定精度は低下してしまう．現実には採点コストを軽減するために，受検者あたりの評価者数は少ないことが多いため，この点は実用上の問題となる．

そこで我々は，論述式試験を対象にこの問題を解決する手法の一つとして，評価者が与える評点データに加えて，回答文の文章情報も加味して能力を推定できるモデルを提案した[7, 16]．このモデルは，自然言語処理分野で広く利用されるトピックモデルを統合した項目反応モデルとして定式化した．具体的には，トピックモデルのひとつである潜在ディリクレ配分法（Latent Dirichlet Allocation）を用いて各回答文のトピック分布（潜在的な話題・意味を表す）を推定し，そのトピック分布を受検者の能力推定値に反映させるようにモデル化を行った．このモデルでは，評価者が与える評点に加えて，

回答文の内容的な特徴も考慮して能力推定がなされるため，既存モデルより高精度な能力測定が可能であり，回答文あたりの評価者数の減少に伴う能力測定精度の低下を緩和できる．

本研究の成果は，電子情報通信学会論文誌と国際会議 AIED に採択された[19,36]．また，国際会議 AIED では Best paper にノミネートされ，国内学会では人工知能学会で受賞をしている[受賞欄 25].

2.5 マルコフ連鎖モンテカルロ法に基づくベイズ推定アルゴリズム

項目反応理論におけるパラメータ推定手法としては，EM アルゴリズムを用いた周辺最尤推定法やニュートンラフソン法による事後確率最大化推定法が広く用いられてきた．一方で，本稿で紹介したような複雑なモデルの場合には，マルコフ連鎖モンテカルロ（Markov Chain Monte-Carlo: MCMC）を用いた期待事後確率推定法が一般に高精度である．項目反応理論における MCMC アルゴリズムとしては，メトロポリスヘイスティングスとギブスサンプリングを組み合わせたアルゴリズムが利用されてきた．このアルゴリズムは単純で実装が容易である反面，目標分布への収束が遅いという問題がある．より効率の良い MCMC アルゴリズムとして，ハミルトニアンモンテカルロ法やそれを発展させた No-U-Turn Sampler（NUT）と呼ばれる手法が提案されている．特に NUT は，Stan と呼ばれるライブラリの整備により，様々な数理モデルに容易に適用できるようになったため，項目反応理論を含む様々なデータ分析・機械学習モデルの推定に近年広く利用されている．

我々の研究では，一般化多相ラッシュモデル[14]とルーブリック評価のためのモデル[5, 10, 16, 18, 29]，評価者特性の時間変動を考慮した時系列モデル[4]で，NUT に基づく MCMC 法を採用している．原論文では Stan コードも公開している．

2.6 評価者特性を考慮した項目反応理論に基づくパフォーマンステストの等化精度評価

現実の評価場面では，複数回の異なるパフォーマンステストの結果を比較するニーズがしばしば生じる．このような場合に項目反応モデルを適用するためには，個々のテスト結果から推定されるモデルパラメータを同一尺度上に位置付ける「等化」が必要となる．一般に，パフォーマンステストの等化を行うためには，テスト間で課題と評価者の一部が共通するように個々のテストを設計する必要がある．このとき，等化の精度は，共通課題や共通評価者の数，各テストにおける受検者の能力特性分布，受検者数・評価者数・課題数などの様々な条件に依存すると考えられる．しかし，これまで，これらの要因が等化精度に与える影響は明らかにされておらず，テストをどのように設計すれば高精度な等化が可能となるかは示されてこなかった．そこで我々は，項目反応モデルをパフォーマンス評価に適用して等化を行う場合に，その精度に影響を与える要因を実験により明らかにし，その結果に基づき，高い等化精度を達成するために必要なテストのデザインについて基準について検討も行なった．

本研究の成果は，国際論文誌 Behavior Research Methods に採択された[13].

2.7 医療系大学間共用試験における実証実験

我々は，パフォーマンス評価のための項目反応理論の実用化・実証実験も推進している．特に，全国の医療系大学の学生が受験する医療系大学間共用試験では，OSCE と呼ばれる実技試験が実施において本技術の実証実験を進めるとともに，全国の医療系大学に向けた講演なども継続的に行っている．

3. 記述式試験の自動採点技術

パフォーマンス評価のための項目反応理論は評価者バイアスを取り除くことによる評価の信頼性改善に寄与するものであった．他方で，主に記述・論述式テストを対象に，人間の評価を代替する技術として，自動採点技術が古くから研究されている．

従来の自動採点技術のアプローチは大きく二つに分類できる．一つは，事前に人手で設計した特徴量（Handcrafted feature）を用いる方法であり，古くから用いられてきたアプローチである．もう一つの方法は，機械学習モデルに単語の系列データを入力し，人手での特徴量設計を行うことなく得点予測を行う方法である．後者の手法は深層学習技術の発展とともに近年特に活発に研究されている．深層学習を用いた自動採点モデルには，様々なモデルが提案されているが，初期の代表的な深層学習自動採点モデルとしては，リカレントニューラルネットワーク（Recurrent Neural Networks: RNN）の一種である Long Short Term Memory （LSTM）と畳み込みニューラルネットワーク（Convolutional Neural Networks: CNN）を組み合わせたモデルが有名である．一方で近年では，BERT に代表される事前学習済みの Transformer モデルを自動採点に用いる方法が一般的である[1]．以降では，深層学習自動採点の性能改善やさまざまな発展課題の解決を目標に，我々が行ってきた研究の成果を概説する．

3.1 項目反応理論を利用した評価者バイアスに頑健な深層学習モデル

深層学習自動採点モデルを利用するためには，事前に収集した大量の採点済み答案データを用いてモデルの学習を行う必要がある．大量の答案の採点作業は一般に多数の評価者で分担して行われるが，そのような場合，個々の答案に与えられる得点が評価者の特性に強く依存してしまう問題が知られている．このような評価者バイアスの影響を受けたデータから自動採点モデルを学習すると，評価者バイアスの影響がモデルにも反映されてしまい，予測性能が著しく低下する．

そこで我々は，評価者特性パラメータを付与した項目反応モデルを自動採点モデルに組み込むことで，評価者バイアスに頑健な深層学習自動採点手法を提案した．本手法は，学習データ中の評価者バイアスの問題に着目した初めての手法であり，様々な自動採点モデルにおいて評価者バイアスに頑健なモデル学習と得点予測を実現できる．

本研究の成果は，電子情報通信学会論文誌，国際論文誌 IEEE Transactions on Learning Technologies，および国際会議 AIED に採択された[7, 9, 32]．また，AIED では Best paper runner-up を受賞し，国内では日本テスト学会でも受賞をしている[受賞欄 21, 22].

3.2 特徴量を組み込んだ深層学習自動採点モデル

人手で作成した特徴量を利用する自動採点手法と深層学習に基づく自動採点手法は独立に研究されることが多いが，これらの二つのアプローチは本来は競合する手法ではなく，それぞれに異なる利点を有している．具体的には，深層学習ベースの手法は語彙の出現パタンに基づいて文全体の意味を分析することで，対象とするデータセットに合わせた特徴量を自動で獲得できるという利点がある．これに対し，特徴量ベース手法では，長年の研究で有効性が検証されてきた高度な特徴量を利用することで，単語の出現パタンだけでは捉えにくい特徴を扱えるという利点がある．

そこで，申請者らは，これらの二つのアプローチを統合した新たなハイブリッド手法を提案した[4]．提案手法は，深層学習モデルで得られる特徴表現ベクトルに人手で設計した文章レベルの特徴量を統合する手法である．本手法は，既存の様々な深層学習自動採点モデルに容易に適用することができ，これまでに開発されてきた有効な特徴量を活用することで，精度を大きく改善できる．

さらに，このアプローチに関連する研究として，文章の論理構造に着目した深層学習自動採点モデルを紹介する．この手法では，論理構造を推定する「論述構造解析」と呼ばれる自然言語処理技術を用いて，回答文の論理構造を推定し，その論理構造を一種の特徴量として深層学習自動採点の枠組みの中で扱えるようにしている．本手法は，文章の論理構造を処理する深層学習モデルと，文章の単語系列を分析する従来の深層学習自動採点モデルの出力を統合して，最終的な得点を予測する構造とし

[1] 直近では GPT4 や Gemini のような大規模言語モデルを用いた Zero-shot/Few-shot での自動採点アプローチの検討が急速に広がっているが，本科研内の成果では事前学習とファインチューニングのアプローチを中心とした研究を行っている．

て設計した．ここで，論理構造を処理する深層学習モデルは，BERT のセルフアテンション機構に工夫を加えたモデルになっている．具体的には，通常のセルフアテンション機構が全ての単語間の関係を参照し合うところを，提案手法では，論理的な関係がある文節内の単語同士でのみ情報を参照し合うように変更することで，論理構造に焦点化したデータ処理を実現している．提案手法は，特に論理構造が重要となる課題において，自動採点の精度を大きく改善できることが実験から示された．

本研究の成果は，国際会議 AIED と COLING に採択された[21, 31]．また，人工知能学会全国大会で優秀賞を，人工知能学会研究会で奨励賞を受賞をしている[受賞欄 5, 10].

3.3 複数の評価観点に基づく自動採点モデル

従来の自動採点研究の多くは，単一の総合得点を予測する手法が主流であった．しかし，現実の評価場面では，評価基準表に基づいて複数観点で採点される場合がある．そこで我々は，総合得点に加えて評価観点別の得点も予測できる自動採点手法を開発した．具体的には，従来の深層学習ベースの自動採点モデルの出力層に多次元型の IRT モデルを組み込んだ構造の手法を提案した．既存の観点別自動採点手法と比べた提案技術の利点は，解釈性に優れる IRT を深層学習自動採点モデルに組み込んでいる点にある．提案手法では，得点予測の根拠を「受検者の能力」と「評価観点の特性値」という二つの要因から解釈できるとともに，複数の評価観点の背後に仮定される能力尺度の解釈も可能となる．ベンチマークデータを用いた実験では，提案手法が，AAAI2021 で発表された当時最高性能の観点別自動採点モデルから，有意には精度を落とすことなく，解釈性の向上に成功したことを示した．

本研究の成果は，電子情報通信学会論文誌と国際会議 COLING に採択された[3, 25]．また，電子情報通信学会の研究会と人工知能学会研究会で受賞をしている[受賞欄 15, 19].

3.4 項目反応理論を用いた複数自動採点機のアンサンブル手法

ここまでに紹介してきた通り，近年では多くの自動採点モデルが提案されており，それぞれに異なった特徴を有している．そこで，本研究では，特徴の異なる多数の自動採点手法を統合するアンサンブルのアプローチによって自動採点の性能の向上を目指した．具体的には，2 章で紹介した評価者特性を考慮した項目反応理論を用いて自動採点モデルのモデル平均を行う手法を提案した．提案手法は，個別の自動採点モデルを一人の評価者とみなして評価者特性を考慮した項目反応モデルを適用することで，それぞれの自動採点モデルの特徴を考慮した統合を行う手法として設計した．実験を通して，提案手法が単体の自動採点モデルや，単純なアンサンブル手法と比べて予測精度を向上できることを示した．

本研究の成果は，電子情報通信学会論文誌と IEEE Transactions on Learning Technologies，および国際会議 AIED に採択された[2, 8, 30]. また，人工知能学会全国大会で優秀賞で受賞をしている[受賞欄 20].

3.5 短答記述式自動採点モデル

以上は，比較的長い回答となる論述式問題を念頭に置いた手法であったが，我々は短答記述式問題を対象とした自動採点モデルの研究も行なっている．以降では，2 つの短答記述式問題自動採点研究の成果を紹介する．

3.5.1 Mixed Format テストを想定した短答記述式自動採点モデル

一つ目の研究では，短答記述式問題が客観式問題を含むテストの一部としてしばしば出題されることに着目している．テストは特定の能力を測定するツールであるため，同一テスト上の短答記述式問題と客観式問題が測定する能力には共通部分が存在すると仮定できる．このことは，同一テスト内の客観式問題から推定される各受検者の能力が短答記述式問題の得点予測の補助情報になりうること

を示唆している．そこで本研究では，客観式問題への正誤データから推定される受検者の能力値を加味できる新たな深層学習自動採点モデルを提案した．具体的には，深層学習自動採点モデルの内部で獲得される回答文の分散表現ベクトル（回答文の特徴を表す固定次元の実数ベクトル）に，客観式問題への正誤データから項目反応理論を用いて推定される受検者の能力値を統合して，回答文の得点を予測するモデルを開発している．

本研究の成果は，教育システム情報論文誌と国際会議 AIED に採択された[12, 33]．また，教育システム情報論文誌では論文賞を，NLP 若手の会では萌芽研究賞を受賞をしている[受賞欄 13, 24].

3.5.2 受検者特徴量を用いた複数問題同時自動採点技術

二つ目の研究では，一つのテストの中に複数の短答記述式問題が出題される場合を想定する．同一テストで複数の短答記述式問題を出題する場合，それらの問題の背後には関心下の測定対象能力が存在すると仮定できる．この仮定のもとで，本研究では，各受検者の複数の問題に対する答案文集合から，その受検者に固有の特徴量を抽出し，得られた受検者特徴量と個々の回答文の分散表現ベクトルを用いて，各回答文の得点予測を行う多入力・他出力型深層学習モデルを提案した．実験の結果，提案手法は従来手法より高精度な自動採点を実現し，さらに，訓練を受けた人間による採点に匹敵する精度で自動採点ができたことが確認された．

本研究の成果は，国際会議 ICALT に採択された[22].

3.6 自動採点技術を用いた IRT の等化手法

ここでは，自動採点技術の応用研究の一つとして，自動採点技術を用いた IRT の等化手法を紹介する．本研究では，異なる受検者集団を対象に同一の記述問題を出題し，それらの回答を受検者集団ごとに異なる評価者集団が採点した場合を想定し，一般化多相ラッシュモデルなどの IRT モデルを利用して同一尺度上で IRT パラメータを推定することを目指す．しかし，2.6 節で述べたように，このような状況では，IRT パラメータの推定値を集団間で直接的に比較できない．この問題を解決する方法の一つとしては，集団間に共通の評価者を配置するデザインで全集団の評点データを収集し，共通の評価者の得点をアンカーとして，IRT パラメータを推定する方法が一般的である．しかし，実際には，共通評価者を用意できない場合もありえる．

そこで本研究では，3.1 節で紹介した「評価者バイアスに頑健な自動採点技術」を応用することで，共通評価者なしに等化を実現することを目指した．評価者バイアスに頑健な自動採点技術は，訓練データ中の評価者バイアスを取り除いて，安定的なモデル訓練を行うために提案したが，この技術の別の特徴として，IRT に基づく受検者の能力値を答案文から訓練データの尺度に沿って予測できるという副次的な利点がある．本研究では，この性質を利用することで，共通評価者なしに等化を実現する手法を開発した．具体的には，この手法では，まず，基準となる集団（基準集団）の得点データに IRT を適用して基準集団に属する受検者の能力値を推定し，それらの受検者の能力値と答案文集合を用いて「能力値を予測する自動採点モデル」を訓練する．次に，訓練された自動採点モデルに，等化対象となる受検者集団の答案分データを入力することで，その受検者集団に対する能力予測値を求める．ここでのポイントは，この能力予測値が基準集団に属する受検者集団の能力尺度に従って予測されているという点にある．最後に，等化対象集団の得点データから推定される能力推定値の尺度を，自動採点モデルで予測した能力予測値の尺度に合わせるように，全体のパラメータの尺度を調整する．これによって，等化対象集団のデータから推定される全ての IRT パラメタ値の尺度を，基準集団から推定されるパラメタ値の尺度に等化することが可能となる．

本研究の成果は，国際会議 IMPS で発表し，現在，国際論文誌 Behavior Research Methods に投稿中（改訂中）である．

4. 問題自動生成

急速な情報化の進展に伴い，様々な情報の中から必要な情報を取捨選択し，内容を正確に理解する読解力がこれまで以上に求められている．読解力の育成方法の一つとして，学習者に多様な文章を読ませ，それに関連する読解問題に取り組ませるアプローチが知られている．しかし，様々な読解対象文に対する多様な読解問題を人手で作成することは時間的・費用的コストが高いという問題がある．

この問題を解決する方法の一つとして，読解問題自動生成技術が近年注目を集めている．読解問題自動生成とは，読解対象文からそれに関連する問題を自動生成する技術であり，教育分野においては読解力の育成・評価を支援する技術の一つとして活用が期待されている．

従来の読解問題自動生成手法では，人手で設計したルールやテンプレートを利用して原文を問題文に変換するアプローチが主流であったが，近年では，深層学習を用いた end-to-end の読解問題自動生成手法が多数提案され，高性能を達成している．

このような問題生成技術を読解力育成のための学習支援として活用する場合，生成できればどのような問題でもいいわけではなく，対象とする学習者の能力に合わせた適切な難易度の問題を出題することが効果的と考えられる．しかし，従来の問題生成手法では，難易度を細かくコントロールした問題生成は直接には実現できなかった．

これらの問題を解決するために本研究では，テスト理論の一つである項目反応理論を用いて問題難易度を定量化し，その難易度値を指定して読解対象文から問題文と答えを大規模言語モデルで自動生成する手法を提案した. IRT は問題の難易度と学習者の能力の関係性をモデル化する理論であるため，IRT を活用することで学習者の能力にあった難易度値の選定が可能となる．本研究で提案する答えと問題の生成手法は，1）読解対象文と IRT に基づく難易度を入力として，読解対象文から答えを抽出する難易度調整可能な答え抽出モデルと，2）読解対象文と答え，IRT に基づく難易度を入力として，問題文を生成する難易度調整可能な問題生成モデルで構成される．難易度調整可能な答え抽出モデルでは基礎モデルに BERT を，難易度調整可能な問題生成モデルでは基礎モデルに Text-to-Text Transfer Transformer（T5）を利用した．また，提案手法を学習するためには，＜読解対象文，問題文，答え，IRT に基づく難易度＞の 4 つ組で構成されるデータセットが必要となるが，読解問題自動生成の従来研究で利用されるベンチマークデータセットは＜読解対象文，問題文，答え＞の 3 つ組で構成されている．そこで本研究では，従来のデータセット中の各問題を性能に差がある多数の質問応答（Question Answering: QA）システムに解かせることで，＜読解対象文，問題文，答え，IRT に基づく難易度＞の 4 つ組で構成されるデータセットを構築する手法の提案も行った．さらに，学習者の能力に合った問題を生成するためには学習者の能力が既知である必要があるが，通常の学習場面ではこの前提は満たされない．そこで，本研究では，テストの出題方式の一つとして知られる適応型テスト（Computerized Adaptive Testing: CAT）の枠組みを利用することで，学習者の能力を効率的に推定しつつ能力に合った難易度の問題を逐次的に提示するアプローチも提案している．

実データ実験の結果，提案手法で生成された問題のうち，おおよそ 9 割は十分に流暢，あるいは許容可能な文章で生成されており，おおよそ 9 割の問題は，読解文に関連した適当な問題となっていることが確認された．さらに，約 9 割の問題がそのままで解答可能，または若干の問題修正で解答可能になることがわかった．また，難易度をばらつかせて生成した問題のうち，人間の評定者に解答可能と判定された問題 30 問を人間の受検者 10 名に回答させ, その正答率と難易度の相関を調べたところ，相関は-0.67 となり , 無相関検定により 1%水準で有意な相関があることが確認された. このことから，難易度を反映させた問題生成が実現できていることが確認された．

本研究の成果は，電子情報通信学会論文誌と ACL の Workshop に採択された[1, 20]．また，国内の複数の学会（電子情報通信学会，人工知能学会，教育システム情報学会）や産学連携に関するイベント（ラーニングイノベーショングランプリ）で多数の賞を受賞した[受賞欄 1, 3, 4, 6, 7, 8, 9, 11, 14, 18].

5. むすび

本資料では，1）パフォーマンス評価のための項目反応理論とその応用技術，2）記述式回答の自動採点技術，3）問題自動生成技術，のそれぞれに関して，基盤研究 S「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」で得られた成果の概要を報告した．それぞれの技術の詳細は，各セクションで示した個別の論文を参照いただきたい．また，下記の資料・情報も適宜参考いただければと存じます．

科研費基盤 S シンポジウム（2023 年 12 月開催）におけるプレゼン資料
https://drive.google.com/file/d/19-GO7tu1bPufAXesBAcfI_eubIw4dkCL/view

関連業績一覧（論文へのリンク付き）
https://sites.google.com/view/utomasaki/KibanS

業績リスト

基盤研究 S「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」の実施期間（2019 年度～2023 年度）のおける査読付きの研究業績および受賞歴を以下に示す．

【査読付き論文誌】

[1] 富川雄斗・鈴木彩香・**宇都雅輝**（2024）項目反応理論に基づく難易度調整可能な読解問題自動生成手法．電子情報通信学会論文誌 D, Vol.J107-D, No.02, pp.53-66.

[2] **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.

[3] 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-D, No.01. pp.47-56.

[4] **Masaki Uto** (2023) A Bayesian Many-Facet Rasch Model with Markov Modeling for Rater Severity Drift. Behavior Research Methods, Springer, Vol.55, 3910-3928. [IF=5.953]

[5] **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-D, No.07, pp.457-469.

[6] Minoru Nakayama, Filippo Sciarrone, Marco Temperini, Masaki Uto (2022) An Item Response Theory Approach to Enhance Peer Assessment Effectiveness in Massive Open Online Courses. International Journal of Distance Education Technologies, Vol.20, No.1, pp.1-19.

[7] **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, Issue 6, pp.763-776. [IF: 4.433]

[8] 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-D, No.11, pp.784-795.

[9] 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-D, No.8, pp.650-662.

[10] **Masaki Uto** (2021) A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. Behaviormetrika, Springer, Vol.48, Issue 2, pp.425-457.

[11] **Masaki Uto** (2021) A review of deep-neural automated essay scoring models. Behaviormetrika, Springer, Vol.48, Issue 2, pp.459-484.

[12] 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. （論文賞受賞）

[13] **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. [IF=5.953]

[14] **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.

[15] **Masaki Uto**, Yoshimitsu Miyazawa, Yoshihiro Kato, Koji Nakajima, Hajime Kuwata (2020) Time- and learner-dependent hidden Markov model for writing process analysis using keystroke log data. International Journal of Artificial Intelligence in Education, Springer, Vol. 30, No.2, pp.271-298. [IF: 4.9]

[16] **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.

[17] **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. [IF: 4.433]

[18] 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.

[19] **宇都雅輝** (2019) 論述式試験における評点データと文章情報を活用した項目反応トピックモデル．電子情報通信学会論文誌 D. Vol.J102, No.8, pp.553-566.

【国際会議発表】

[20] **Masaki Uto**, Yuto Tomikawa, Ayaka Suzuki (2023) Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory.18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), Association for Computational Linguistics (ACL), pp.119-129.

[21] Misato Yamaura, Itsuki Fukuda, **Masaki Uto** (2023) Neural automated essay scoring considering logical structure. 24th International Conference on Artificial Intelligence in Education (AIED), pp.267-278. [Accepted as full paper, full paper acceptance rate= 21.1%, CORE-Rank=A]

[22] **Masaki Uto** (2023) Neural Automated Short-Answer Grading Considering Examinee-Specific Features. 23rd IEEE International Conference on Advanced Learning Technologies (ICALT), pp.336-338. [Accepted as short paper, CORE-Rank=B]

[23] Kota Aramaki, **Masaki Uto** (2023) Linking method for writing tests using item response theory and automated essay scoring. International Meeting of the Psychometric Society (IMPS).

[24] Minoru Nakayama, **Masaki Uto**, Satoru Kikuchi, Hiroh Yamamoto (2023) Feasibility of Prediction of Student's Characteristics using Texts of Essays Written during a Fully Online Course. 27th International Conference on Information Visualisation (IV), pp.204-209.

[25] 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), [Accepted as full paper, full paper acceptance rate= 24.2%, CORE Rank=A]

[26] Minoru Nakayama, Satoru Kikuchi, **Masaki Uto**, Hiroh Yamamoto (2022) Evaluation of Essays and Comments for Developing Critical Thinking Ability during a University course. The Workshop on Psychology Learning Technology (PLS).

[27] Minoru Nakayama, Filippo Sciarrone, Marco Temperini, **Masaki Uto** (2022) Evaluation of Programming Skills via Peer Assessment and IRT Estimation Techniques. 20th International Conference on Information Technology Based Higher Education and Training (ITHET), pp.1-8.

[28] Minoru Nakayama, **Masaki Uto**, Marco Temperini, Filippo Scarrone (2021) Estimating Ability of Programming Skills using IRT based Peer Assessments. 19th International Conference on Information Technology Based Higher Education and Training (ITHET), pp.1-6.

[29] **Masaki Uto** (2021) A Multidimensional Item Response Theory Model for Rubric-based Writing Assessment. International Conference on Artificial Intelligence in Education (AIED), Lecture Notes in Computer Science, vol.12748, pp.420–432. [Accepted as full paper, full paper acceptance rate= 24%, CORE Rank=A] <Best paper award nominee>

[30] 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), Lecture Notes in Computer Science, vol.12749, pp.54–59. [Accepted as short paper, CORE Rank=A]

[31] **Masaki Uto**, Yikuan Xie, Maomi Ueno (2020) Neural Automated Essay Scoring Incorporating Handcrafted Features. Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp.6077-6088. [Accepted as full paper, CORE Rank=A]

[32] **Masaki Uto**, Masashi Okano (2020) Robust neural automated essay scoring using item response theory. International Conference on Artificial Intelligence in Education (AIED), Lecture Notes in Computer Science, vol 12164, pp.549-561. [Accepted as full paper, full paper acceptance rate= 26.6%, CORE Rank=A] <Best paper runner-up award>

[33] **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), Lecture Notes in Computer Science, vol 12164, pp.334-339. [Accepted as short paper, CORE Rank=A]

[34] Minoru Nakayama, Filippo Sciarrone, **Masaki Uto**, Marco Temperini (2020) Impact of the number of peers on a mutual assessment as learner's performance in a simulated MOOC environment using the IRT model. 24th International Conference Information Visualization (IV). pp. 483-487.

[35] Minoru Nakayama, Filippo Sciarrone, **Masaki Uto**, Marco Temperini (2020) Estimating student's performance based on item response theory in a MOOC environment with peer assessment. International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL), Advances in Intelligent Systems and Computing, Springer, vol 1236, pp. 25-35.

[36] **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), pp. 494-506. [Accepted as full paper, full paper acceptance rate= 25%, CORE Rank=A]

【受賞】

1. 教育システム情報学会 学生研究発表会 優秀発表賞（2024）後藤照佳・富川雄斗・宇都雅輝「問題と模範解答を同時に生成する難易度調整機能付き読解問題自動生成手法」（受賞者：後藤照佳）
2. 日本行動計量学会 優秀賞（林知己夫賞）（2023）受賞者：宇都雅輝
3. 電子情報通信学会教育工学研究会 研究奨励賞（2023）後藤照佳・富川雄斗・宇都雅輝「問題と模範解答の同時生成機構を持つ難易度調整可能な読解問題自動生成手法」（受賞者：後藤照佳）
4. 教育システム情報学会 第 48 回全国大会 大会奨励賞（2023）富川雄斗・宇都雅輝「読解対象文の難易度を考慮した読解問題自動生成手法」（受賞者：富川雄斗）
5. 人工知能学会 全国大会優秀賞（2023）山浦美里・福田樹・宇都雅輝「論述構造解析を用いたニューラル小論文自動採点手法の提案」（受賞者：山浦美里・宇都雅輝）
6. ラーニングイノベーショングランプリ 2023 優秀ラーニングイノベーション賞（2023）宇都研究室 問題生成グループ「深層学習と項目反応理論を用いた難易度調整可能な読解問題自動生成」（グループ構成員：富川雄斗・鈴木彩香・後藤照佳・宇都雅輝）
7. ラーニングイノベーショングランプリ 2023 特別賞：UMU ラーニングテクノロジー賞（2023）宇都研究室 問題生成グループ「深層学習と項目反応理論を用いた難易度調整可能な読解問題自動生成」（グループ構成員：富川雄斗・鈴木彩香・後藤照佳・宇都雅輝）
8. 教育システム情報学会 産学連携奨励賞（2023）富川雄斗・宇都雅輝「読解対象文の難易度を考慮した読解問題自動生成手法」（受賞者：電気通信大学 宇都研究室）
9. 人工知能学会 研究会優秀賞（2023）鈴木彩香・宇都雅輝「項目反応理論と深層学習を用いた難易度調節可能な読解問題自動生成手法」（受賞者：鈴木彩香・宇都雅輝）
10. 人工知能学会 先進的学習科学と工学研究会 令和 4 年度 若手奨励賞（2023）山浦美里・福田樹・宇都雅輝「論述構造解析を組み込んだニューラル小論文自動採点手法」（受賞者：山浦美里）
11. 教育システム情報学会 学生研究発表会 優秀発表賞（2023）富川雄斗・宇都雅輝「読解対象文の難易度を考慮した読解問題自動生成手法」（受賞者：富川雄斗）
12. 教育システム情報学会 学生研究発表会 支部長賞（2023）高橋祐斗・宇都雅輝「アンサンブル法に基づく深層学習自動採点の不確かさ推定」（受賞者：高橋祐斗）
13. 教育システム情報学会 論文賞（2022）内田優斗・宇都雅輝「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」（受賞者：内田優斗・宇都雅輝）
14. 教育システム情報学会 第 47 回全国大会 大会奨励賞 （2022）鈴木彩香・宇都雅輝「深層学習を用いた難易度調整機能付き読解問題自動生成手法」（受賞者：鈴木彩香）

15. 人工知能学会 先進的学習科学と工学研究会 令和 3 年度 若手奨励賞（2022）柴田拓海・宇都雅輝「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」（受賞者：柴田拓海）
16. 日本テスト学会第 19 回大会 大会発表賞（2022）岡野将士・宇都雅輝「深層学習自動採点技術を組み込んだ一般化多相ラッシュモデル」（受賞者：岡野将士・宇都雅輝）
17. 教育システム情報学会 学生研究発表会 優秀発表賞（2022）林真由・宇都雅輝「評価者特性の時間変動を考慮した項目反応モデル」（受賞者：林真由）
18. 教育システム情報学会 学生研究発表会 支部長賞（2022）鈴木彩香・宇都雅輝「難易度調整機能を持つ GPT-2 に基づく読解問題自動生成手法」（受賞者：鈴木彩香）
19. 電子情報通信学会教育工学研究会 研究奨励賞（2021）柴田拓海・宇都雅輝「深層学習と多次元項目反応理論を用いた複数観点同時自動採点手法の開発」（受賞者：柴田拓海）
20. 人工知能学会全国大会優秀賞（2021）青見樹・堤瑛美子・宇都雅輝・植野真臣「項目反応理論を用いた自動採点モデルの統合手法」（受賞者：青見樹・堤瑛美子・宇都雅輝・植野真臣）
21. Best paper runner-up award, International Conference on Artificial Intelligence in Education (AIED) (2020) Masaki Uto, Masashi Okano: Robust neural automated essay scoring using item response theory. [Full paper acceptance rate= 26.6%, CORE Rank=A] <<Certification>>
22. 人工知能学会 先進的学習科学と工学研究会 令和元年度 若手奨励賞（2020）岡野将士・宇都雅輝「評価者バイアスに頑健な小論文自動採点手法」（受賞者：岡野将士）
23. 日本行動計量学会 奨励賞（肥田野直・水野欽司賞）（2019）受賞者：宇都雅輝
24. NLP (言語処理) 若手の会第 14 回シンポジウム 萌芽研究賞（2019）内田優斗・宇都雅輝「受験者の解答履歴データを組み込んだ短答式問題自動採点手法」（受賞者：内田優斗・宇都雅輝）
25. 人工知能学会 研究会優秀賞（2019）宇都雅輝「レイティングデータとテキスト情報を用いて受験者の能力を推定する項目反応トピックモデルの提案」（受賞者：宇都雅輝）
26. 人工知能学会 先進的学習科学と工学研究会 平成 30 年度 若手奨励賞（2019）八木嵩大・宇都雅輝「パフォーマンス評価における多次元段階反応モデルの提案と評価」（受賞者：八木嵩大）

東京医科歯科大学での３D シミュレータによる実技訓練及び OSCE における実証実験

荒木孝二[1]，鶴田　潤[1]

[1]東京医科歯科大学　統合教育機構

1. はじめに

日本の歯科大学の学生教育は、基本的に臨床実習のための技術教育と患者とのコミュニケーション教育が行なわれている。そのために臨床実習開始前に臨床手技と医療面接・感染予防・安全対策における能力の向上が求められている。これらの教育成果の確認として臨床実習開始前の共用試験が実施されており、全 29 歯科大学が参加している。この共用試験の合格基準に到達した学生が、各大学で実施する臨床実習に参加できる。共用試験には知識を確認する CBT と、基本的臨床技能と態度を確認するパフォーマンステストとしての OSCE の 2 つがある。共用試験は歯科医師国家試験のように全国同時に行なわれるのではなく、各大学のカリキュラムに合わせて単独で行なわれる。そのため、OSCE 実施には、実施大学の教職員の相当数が関与する準備と実施体制、及び全国から派遣される実施大学以外の評価者が必要となっている。パフォーマンステストとしての OSCE に対して、いかに効率よく厳格公正に実施及び信頼性・妥当性の高い評価が行えるのかが喫緊の課題となっている。

OSCE の実施時に、受験生が行なった課題に対してビデオ撮影を行ない、評価者が別室あるいは後日実施することが考えられているが、積極的には行なわれていない。

また、臨床手技の訓練装置として近年諸外国急速に普及しつつある 3D バーチャルリアリティシミュレータに関する我が国への導入はほとんど進んでいない。従来から行なわれているマネキンヘッド型のシミュレータとの比較効果の研究はほとんど行なわれていない。

東京医科歯科大学での実証実験

東京医科歯科大学歯学部には学生が臨床技能を自習できるスキルスラボラトリーが設置されており、従来型のマネキンヘッドを用いて顎模型を使用できるシミュレータとすべての臨床手技を PC 画面上で実施できる触覚型 3D バーチャルリアリティシミュレータ 2 台がある。そこで、3D バーチャルリアリティシミュレータとマネキンヘッド型のシミュレータとの歯質切削訓練の比較について行なった。また、臨床実習終了時に実施している OSCE 医療面接の課題実施時に複数方向からビデオ撮影をおこない、後日複数の評価者が実施した評価項目についての信頼性・妥当性の分析と項目反応理論への応用に関する実証実験を行なった。

2.1　3D バーチャルリアリティシミュレータの歯質切削訓練への効果

歯科臨床教育にマネキンヘッドを用いた模擬診療的なシミュレータ（以下、従来型シミュレータ）に加え、診療手技を画面上で行う触覚型 3D バーチャルリアリティシミュレータ（以下、3D 触覚シミュレータ）が導入されてきた。3D 触覚シミュレータでは、3D モニター上で切削感を伴う訓練が可能であり、客観的評価の観点でもデジタル教育機器の有用性が報告されており臨床教育での活用が期待される[1] [2]。一方、教育現場では、従来型シミュレータでの指導が大半を占めており、教育導入には両シミュレータの特性の違いを明らかにすることが重要と考える。本研究では，同一学生が両シミュレータで作成した成果物に対する同一評価者の評価を比較し、シミュレータの違いが成果物およびその評価にどのような影響を及ぼすかを検証した。

2019 年度東京医科歯科大学歯学部歯学科 6 年生のうち、実験参加への同意が得られた 30 名を対象とした。3D 触覚シミュレータとして、Simodont®（NISSIN DENTAL PRODUCTS EUROPE B.V.）を使用した。従来型シミュレータとして、Clinsim（クリンシム 臨床シミュレーションシステム、株式会社モリタ製作所）を用い、人工歯および顎模型（株式会社ニッシン）を装着し使用した。

対象者には支台歯形成を行う前に、Simodont®と Clinsim それぞれで下顎右側第一大臼歯の全部鋳造冠の支台歯形成を行った。形成済人工歯を提示し目標とするよう指示し、形成条件は、1）咬合面概形は縮小型、2）マージン形態はライトシャンファー、3）テーパー角度は片側で 2～5°、4）咬合面削除量は 1.2～1.5ｍｍとした。各シミュレータでの制限時間は１時間とし時間内なら何本でも形成して良いこととし、対象者自身の判断で最もよく出来たと思う成果物を１本ずつ指定してもらい、評価に使用した。

成果物の評価方法は、大学教員歴・臨床経験歴 10 年以上の 3 名の評価者が成果物の評価を行った。評価項目は①咬合面概形②マージン形態③形成面のなめらかさ④テーパーの状態⑤形成量とし、それぞれ５段階レーティングスケールで評価した。また、概略点を 10 点満点で採点した。また、Simodont®では成果物の STL データ、Clinsim では成果物の規格写真（水平面・矢状面）で得られた計測実測値を評価した。さらに、対象者の Simodont®への主観的評価検討のために、実験後にアンケート調査を行った。「Simodont®は形成練習において有効と感じますか」の問いへの 5 段階レーティングスケール、Clinsim と比較して気が付いた点について自由記載で回答を得た。

Clinsim 成果物評価点では、全項目において、評価者 3、評価者 1、評価者 2 の順に大きな値を示した。級内相関係数は 0.896 であり、3 人の評価者間信頼性は高い結果となった。Simodont®と Clinsim の成果物評価において、②マージン形態、⑤形成量の 2 項目で有意差を認めた($P<0.05$)。①～⑤すべての評価項目で評価者間の採点結果に有意な差を認めた($P<0.05$)。シミュレータと評価者の交互作用を認め両シミュレータでバラつきがみられたのが①咬合面概形、⑤形成量、概略評価であった。近心および遠心のテーパー角度の実測値については、Clinsim 成果物が有意に大きな値を示した($P<0.05$)。

実験後のアンケートの「Simodont®は形成練習に有効と感じますか」の問いに対し、18 名（60%）が「有効」あるいは「それなりに有効」と答えた。従来型との違いとして、「切削感」「対象の観察方向の自由度」「距離感、遠近感」「フィンガーレストの有無」が自由記載に挙げられた。

Clinsim 成果物と比較して Simodont®成果物での評価点が低くなった。対象者の切削技術が一定と仮定したとき、この差はシミュレータの違いにより生じたものと考えられる。このシミュレータの違いが与える影響は、・対象者の切削技能に与える影響と、・評価者の評価過程に与える影響に大別されると考えた。今回これら 2 つの要素を分離して考察するために、線形混合モデルを用い、固定効果として、シミュレータと評価者、およびその交互作用項を説明変数として分析を行った。その結果、シミュレータの違いが対象者の切削技能と有意な関連を認めたのは、マージン形態と形成量であった。対象者の半数以上が両シミュレータの切削感の違いを指摘しており、Simodont®では、フィンガーレストが置けず切削しにくいとの記載もあった。切削感の違いが、対象者の切削技能に影響していると考えられた。また立体映像の認知能力に関しては個人差があることが報告されているが、本実験でも距離感・遠近感がつかみにくいことを指摘しており、立体映像の認知能力が、切削技能に影響していると考えられた。また、Simodont®の特徴として、切削対象歯の観察方向の自由度があり、切削対象歯をモニター上で好きな方向から観察ができる。自由記載では、10 名が実際では解剖学的に不可能な方向から観察できることを指摘していた。

本研究では、両シミュレータ成果物のテーパー角度実測値が、頬舌面では差を認めなかったのに対し、近遠心面では Clinsim 成果物が有意に大きな値を示した。マネキン実習における支台歯形成のテーパー角度についてはこれまで報告があるが、今回設定した目標値よりも大きな実測値が出ている実験が多く、バラつきのある結果となっている[3] [4]。切削対象歯の観察方向の自由度が、手指感覚の違いや立体映像の認知能力の差によって生じる操作性の不利を補ったためと考えられる。

Clinsim 成果物に対する 3 人の評価者による評価結果について、評価者間の級内相関係数は 0.896 であり、従来型シミュレータでの評価者評価の信頼性の高さを示した。一方、評価者間の採点差はすべての評価項目で Clinsim が高い結果となったが、両シミュレータ成果物に対する 3 人の評価点については、6 項目中表面の滑らかさ以外の 5 つの評価項目で採点の上下動があり、シミュレータの違いがそれぞれの評価者に異なる影響を与えていることが考えられた。線形混合モデルで、シミュレータと評価者の交互作用を認めた評価項目は、咬合面概形、切削量、概略点であった。その他の項目、マージン形態、表面のなめらかさ、テーパーが局所に着目する評価項目である一方、交互作用を認めた 3 項目は、評価対象歯の 3 次元的俯瞰が必要な項目であり、評価者間の立体映像の認知様式の違いが評価の違いの交互作用として影響していると考えられた。以上より、Simodont®の特性が評価者の評価過程に影響していることが示唆された。

3D 触覚シミュレータは、対象となるモニター上の画像再構築出来る点は、今後の歯科臨床教育への応用を考える上で特筆すべき点である。しかし、患者診療と同様に実空間で実物に触れられるという点では、従来型シミュレータに劣る点もある。患者診療で切削診療のための準備教育としてのシミュレータの利用を考えると、3D 触覚シミュレータと従来型シミュレータの役割は互いに代替できるものではなく、それぞれの特長を活かした教育プログラムの構築が重要であると思われる。

2.2　OSCE における撮影画像での評価の信頼性・妥当性の分析と項目反応理論の応用に関する実証実験

医学・歯学教育においては、知識領域における教育評価と並び、診療技能や臨床現場での態度に関する教育が重要となっており、その学修者評価については、学習者の行動を直接観察した記録をもとに評価を行う客観的臨床能力試験(OSCE: Objective structured clinical examinations)が行われる。OSCE では、一定の条件下で、学習者（被評価者）が臨床的課題（医療面接、臨床手技など）について実技を行う試験であり、その評価においては、評価者がそれらの行動を直接観察し、関連する複数の項目に関する評価記録を行う。被評価者が行う一連の行動は時系列的に連続しているものであり、評価を行う一定の時間内においては、その行動をその時点で確認する必要がある。このような状況において、評価としての信頼性を高めるために、個人の評価結果の依存度を解消するために複数評価者の配置、評価のばらつきを少なくするための評価項目の設定、評価者トレーニングの実施、複数評価者間における評価基準の共有のためのルーブリックの作成、などの方策がとられるが、その限界として、評価そのものが評価者個人に委ねられていることは残る。この点について、評価者と各評価項目に関する特性を考慮した評価シミュレーションの実施・実証実験を行うこととした。方法として、2021 年に東京医科歯科大学歯学科６年に行われた OSCE 形式での臨床実習後医療面接試験の再確認資料として撮影されたビデオについて、データの 2 次利用について同意の得られた 30 本について、OSCE 評価者経験のある 5 名の歯科医師である教員によって、29 項目の個別評価と総合評価の合わせた 30 項目に関する評価を行った。ビデオについては、PC でビデオファイルして編集を行い 1～30 のビデオ番号を付し、各評価者に、タブレット端末でビデオファイルを配布し、紙媒体でルーブリック表、30 項目

を含む評価シートを配布した。5 名の評価者に対しては互いに評価に関する意見交換を行わないように指示を行い、それぞれの評価が可能な時間において、ビデオ視聴による評価を行うこととした。5 名の評価者について、2 名が 30 名（1～30）全員分の評価、3 名が 10 名ずつ（1～10、11～20、21～30）の評価を行った。評価段階については、4 点満点とし、(4：優れている、3：良い、2：許容できる、1：不満足）とした。30 項目の評価項目は次の項目とした。挨拶、自己紹介、患者確認、説明と同意、主訴の確認、部位の確認、時期の確認、現症（自発痛、腫脹、発赤など）の確認、誘発痛の確認（冷痛、温痛など)、誘発された痛みの性質と重症度の確認、通院歴の確認、主訴に対する投薬の有無の確認、受療行動の確認、症状変化の確認、麻酔および抜歯の既往歴とその際の異常の確認、全身疾患の有無の確認、疾患に対する情報の確認（歯科治療に対する潜在的なリスクが特定できるかどうか)、アレルギーの確認（食物、薬物)、主訴の要約と確認、言い忘れの有無の確認、清潔操作、患者理解に対する行動（意思確認、通院状況、経済状況など、会話のスムーズさ、会話の要点の適切性（会話の要点の判断が患者の状況に応じて適切であったかどうか)、アイコンタクト、間、傾聴、オープン・クエスチョン、敬語、専門用語使用への配慮、患者との関係構築、患者の懸念への配慮、総合評価。

これらの評価結果に対して、IRT（項目反応理論：item response theory）を適用することで、評価者や評価項目の特性を調べ、評価者や評価項目の特性の影響を考慮しながら、受験者の能力を正確に推定する検証を行った。評価者の特性として、厳しさ（他評価者との比較)、評価の一貫性、評価得点の範囲（得点のばらつき）などがあり、評価項目の特性として、難易度、識別力、評価尺度があるところで、IRT を適用することで被評価者の能力を評価するシミュレーション、実データによる実証を行ったものである。実験においては、GMFRM（Generalized the many-facet Rasch model）をもとに検証を行い、結果として、提案モデルの有効性が実証された。評価者の厳しさや評価尺度が評価項目ごとに異なることが明らかになり、提案モデルについては、これら 2 つの特性の影響を考慮しながら被検者の能力を推定するため、従来モデルよりも適合性が向上し、能力測定精度が向上したことが明らかとなった。今回の実証における限界条件としては、そのデータサイズが小さく、パラメータ推定制度に影響を与えている可能性が挙げられる。また、評価者の特性として考えられるハロー効果、評価にかけられる時間による評価者パラメータドリフトなどの影響も考えられる。

2. むすび

医学・歯学教育においては、そのたゆまぬ質の向上が社会から求められており、学修課程修了時に学修者が修得しているコンピテンシーを正しく評価することが必要となっている。学修者評価については、臨床技能を模擬的な状況で行い評価を行う OSCE はじめ、実現場での Workplace Based Assessment（WpBA）など、評価者の直接的な観察による評価活動が重要となるところ、それら評価活動の信頼性がより強く求められる。直接的な観察活動による情報収集の利点としては、学修者のパフォーマンスを臨床活動として観察することにより、実際の医師・歯科医師に必要とされる能力を評価できる点にあるが、一方で、その評価判断が評価者個人に委ねられることとなり、その一貫性、再現性などについての限界が認められることも事実である。この点について、パフォーマンス評価活動として、大人数を一度に一律の環境下で評価する OSCE については、評価活動に関わる条件（直接観察、評価者、評価項目など）を検討することにより、より精度の高い評価を行うことができる可能性があると考えた。今回、項目反応理論を適応するための検証として、能力測定の精度向上に対しての GMFRM の有効性が確認できたことは､現在の OSCE の評価方法の新たな選択肢を提案するものと考えられる。

東南アジア、英国においては、臨床現場での能力を重視するコンピテンシー基盤型教育の実現の一環として、OSCE の活用からシフトし、臨床症例を再現する臨床シミュレーション教育（ハプティクス

ＶＲシミュレータの活用）が盛んとなってきている。一方で、臨床技能の学修環境として一律の条件を適応できるこれらハプティクスＶＲシミュレータについて、そのパフォーマンス評価については、従来の模型歯切削物の評価と同じ方法や単一設定基準に対する数値評価にとどまっており、評価活動の改善、標準化が求められる領域である。今回得られた OSCE 医療面接での新たな評価方法が確立をもって、それらをコンピテンシー評価基準への適用することで、臨床能力評価に対して、より信頼性の厚い評価活動を行うことができる可能性も高いと考える。

プログラム修了時の成果が求められる時代においては、学修者（学部学生、研修医など）がその学修成果として実施的な臨床能力を修得したかを示す記録については、学修者の質評価のエビデンスとして、より厳格なレベルで求められることとなる。今回得た成果は、そのニーズに対するパフォーマンス評価の可能性を示すものであると考える。

引用文献

[1] de Boer IR, Lagerweij MD, Wesselink PR, Vervoorn JM. （2019）The effect of variations in force feedback in a virtual reality environment on the performance and satisfaction of dental students. Simulat Healthc、14(3)、169-174.

[2] Murbay S, Neelakantan P, Chang JWW, Yeung S.（2020） Evaluation of the introduction of a dental virtual simulator on the perfor mance of undergraduate dental students in the preclinical operative dentistry course. Eur J Dent Educ、24(1)、5-16.

[3] Al-Omari WM, Al-Wahadni AM.（2004）Convergence angle, occlusal reduction, and finish line depth of full-crown preparations made by dental students. Quintessence Int、35()、287-293.

[4] Alhazmi M, El-Mowafy O, Zahran MH, Uctasli S, Alkumru H, Nada K.（2013）Angle of convergence of posterior crown preparations made by predoctoral dental students. J Dent Educ、77(9)、1118-1121.

業績リスト

【査読付き論文誌】

1. Tsuruta J.(2023) Simulator education in Japanese dental education. European journal of dental education : official journal of the Association for Dental Education in Europe. December　DOI:10.1111/eje.12983
2. Akitaka Hattori, Ken-Ichi Tonami, Jun Tsuruta, Masayuki Hideshima, Yasuyuki Kimura, Hiroshi Nitta, Kouji Araki (2022) Effect of haptic 3D Virtual reality dental training simulator on assessment of tooth preparation. Journal of Dental Sciences, Vol.14, No.1. pp. 514-520
3. Numasawa Mitsuyuki, Nawa Nobutoshi, Funakoshi Yu, Noritake Kanako, Tsuruta Jun, Kawakami Chiharu, Nakagawa Mina, Yamaguchi Kumiko, Akita Keiichi(2021) A mixed methods study on the readiness of dental, medical, and nursing students for interprofessional learning PLOS ONE Vol 16,No.7: e0255086.
4. Le S. H. Son, Tonami K, Umemori S, Nguyen LT-B, Ngo LT-Q, Araki K, Nitta H.(2020）Relationship between preoperative dental anxiety and short-term inflammatory response following oral surgery. Australian Dental Journal, Vol.66, No.1. pp. 13-19

【国際会議発表】

1. Jun Tsuruta, Atsuhiro Kinoshita(2023), Implementing New Dental Curriculum for Active-Learning, SEAADE 34th Annual Scientific Conference
2. Mina Nakagawa, Kumiko Yamaguchi, Mitsuyuki Numasawa, Kanako Noritake, Eriko Okada, Eiji Kaneko, Janelle Moross, Jun Tsuruta, Keiichi Akita, Masanaga Yamawaki(2023). Patients as teachers: a novel IPE curriculum . AMEE2023
3. Mitsuyuki Numasawa, Nobutoshi Nawa, Kumiko Yamaguchi, Kanako Noritake, Jun Tsuruta, Mina Nakagawa(2022). Comparison of readiness for interprofessional learning among medical, dental, and nursing students before the start of clinical practice. AMEE 2022
4. Mina Nakagawa, Kumiko Yamaguchi, Mitsuyuki Numasawa, Kanako Noritake, Janelle Moross, Jun Tsuruta(2022). Remote interprofessional learning during the COVID-19 pandemic for younger undergraduate students' early exposure to medicine
5. Jun Tsuruta, Ken-ichi Tonami, Kanako Noritake, Kumiko Yamaguchi, Mina Nakagawa(2021). New approach for IPE for dental and medical clinical students with restriction of educational setting under COVID-19 pandemic. Association for Dental Education in Europe, Strasbourg Online Meeting 2021
6. Ken-ichi Tonami, Sachi Umemori, Yasuyuki Kimura, Kanako Noritake, Kouji Araki, Hioshi Nitta (2021) Effects of Online Education on Students 'Self-Reflection about Inter-Personal Relationship. International Association of Dental Research